

All or nothing at Fermilab

These days, it takes a brave laboratory to hitch its future to the construction of a brand-new particle accelerator — but Fermilab has elected to do just that.

It is no secret that high-energy physics faces some immense long-term challenges. Progress relies on the study of particles' behaviour at ever-higher energies, which requires larger machines, and more money to build them. The appetite of governments for such investments has waned in recent years.

These issues are well illustrated by the predicament of the Fermi National Accelerator Laboratory near Chicago. As reported on page 728, Fermilab, the last remaining laboratory in the United States devoted to high-energy physics, is making a daring push to host a new machine, the International Linear Collider (ILC).

The ILC is an electron-positron accelerator that promises to provide further data on the Higgs boson and other particles that physicists hope to discover at CERN, Europe's premier high-energy laboratory, outside Geneva. If Fermilab builds the ILC, it will regain its position at the forefront of international science; failure could lead to staff reductions and intellectual atrophy. But landing the machine won't be easy: Fermilab will need the assurance of large-scale financial support from the US Department of Energy, and the backing of the international physics community.

By choice or necessity, other high-energy physics centres around the world are ensuring that they will not have to rely on the construction of new accelerators. The Stanford Linear Accelerator Center in California, until recently the only other US lab largely devoted to high-energy physics, is reinventing itself as a 'photon science' centre — researchers will use X-rays from its accelerators to study everything from materials to proteins. Similarly, DESY, Germany's primary high-energy physics facility, near Hamburg, is building a billion-euro 'free electron' laser. And the Japanese high-energy physics laboratory KEK is building a neutron source for materials

research and structural biology. They are all retaining high-energy physicists, but are not staking their fate on the next big machine.

That leaves CERN as the only other lab left as dependent on high-energy physics as Fermilab. CERN's immediate outlook is bright, with the Large Hadron Collider (LHC) due to enter service in 2008. But its secondary activities have been depleted by cost overruns on LHC construction, and in the very long run — after LHC's scope is exhausted — its future is as uncertain as Fermilab's.

High-energy physics has always drawn its focus from places such as Fermilab and CERN. They have served as a location for machines, a gathering point for the community, and a base for political support. Today, there are signs that this is changing. The Internet and the Grid — a powerful distributed-computing system now well under development —

are making it easier for university researchers to work on the latest data remotely. Physicists who travel to CERN, for example, won't have access to much more information from the LHC than those back at the universities.

"Other high-energy physics labs are ensuring that they will not have to rely on the construction of new accelerators."

In this interconnected world, the role of the larger centres will evolve. Fermilab's stand in seeking to retain its identity by hosting the linear collider is a bold one. Physicists are at the beginning of an exhaustive international consultation to determine the best design and site for the collider, and the lab will gain a significant advantage by moving aggressively early in the game. But, ultimately, it will require a herculean effort to marshal the political support needed to build the machine, at Fermilab or anywhere else. ■

Kyoto — the sequel

Better knowledge of the carbon cycle could provide a basis for future climate-change agreements.

The Kyoto Protocol has only just come into force, but its 2012 expiration date already looms large on the horizon. A successor agreement faces some formidable political obstacles. But negotiators should put these to one side for now and ask themselves this: what kind of agreement is best justified by our current knowledge of climate science?

Negotiations on Kyoto II are due to begin at the eleventh meeting of the parties to the United Nations Framework Convention on Climate Change in Montreal in November. Climate science, and the tools that researchers can use to detect global change, have advanced significantly since 1997, when the Kyoto Protocol was negotiated.

These improvements must be considered when laying the basis for what comes after 2012.

One important advance has been the study of global carbon cycles — the large-scale carbon fluxes between the biosphere, the atmosphere and the oceans. Siberia, for example, is likely to undergo significant changes in vegetation, wildfire frequency and permafrost abundance, which all have a substantial impact on global carbon transfers (see page 732). Similar changes are being observed in other terrestrial ecosystems, including tropical rainforests and deserts.

Meanwhile, better satellite sensors, in combination with other remote-sensing techniques, provide ever more accurate data about biomass changes on the ground. The Siberian work has demonstrated that it is possible, in principle, to reliably determine carbon flux to and from Earth's main carbon reservoirs.

In a few years' time, a comprehensive, global carbon accounting system, incorporating natural carbon fluxes between the ground, the oceans and the atmosphere, as well as anthropogenic emissions,



could be in place. Global carbon accounting, if it can be shown to work, would greatly reduce uncertainty about changes in atmospheric greenhouse-gas levels. And our understanding of climate sensitivity — the response of the climate to changing greenhouse-gas levels — is slowly improving (see, for example, J. M. Murphy *et al.* *Nature* **430**, 768–772; 2004).

All this could strengthen the case for a Kyoto II agreement that would incorporate a target for the parameter of central importance in determining the course of climate change: the concentration of greenhouse gases in the atmosphere. A new agreement could have such a target at its core, advocates suggest. Such a deal would also include emissions targets, but would be flexible enough to allow changes in these secondary targets as more evidence comes to light, or if natural carbon cycles don't behave as expected.

At their summit meeting in March, European Union (EU) leaders announced a goal of keeping global temperature rise below 2 °C over pre-industrial levels. As temperatures have already increased by around 0.7 °C, this would limit future warming to 1.3 °C. The EU has also tentatively suggested that a 15–30% cut in greenhouse-gas emissions should be sought by 2020.

But targets based on temperature or on emission ceilings still suffer credibility problems. With climate sensitivity still under debate, any given temperature target is necessarily associated with a broad range of possible emissions levels. Likewise, any emission target corresponds to a very wide range of temperature-change scenarios.

Focusing on greenhouse-gas concentrations could help policy-makers steer clear of this uncertainty trap. An interim target of, say, 400 parts per million of carbon dioxide by 2030 might attract widespread political support. Given advances in global carbon accounting, it would also be more manageable and backed by more solid science than the other options. Such a target would have to be supported by goals for greenhouse-gas emissions, but these could be open to some modification in response to real-world observations and advances in scientific knowledge.

The Intergovernmental Panel on Climate Change (IPCC) — the scientific branch of the convention — is already compiling the latest advances in the science of climate change for its next assessment report, due in 2007. The challenge is to incorporate this knowledge into a viable, fair and effective new agreement.

Given the trenchantly negative approach of the current US administration (which will itself expire in 2008), it falls to the EU and developing countries to ensure that the Montreal meeting makes progress towards drafting a new agreement. Their negotiators, backed by the IPCC's findings and the climate-research community, must be decisive in translating the available scientific knowledge into a workable framework for the treaty that will succeed Kyoto. ■

"A Kyoto II agreement could incorporate a target for the concentration of greenhouse gases in the atmosphere."

Talent worth nurturing

More should be done to draw people with disabilities into scientific careers.

As a News Feature in last week's *Nature* amply demonstrated, people with disabilities have a considerable array of talents to bring into science (see *Nature* **435**, 552–554; 2005). But they face an array of obstacles in pursuing scientific careers. According to specialists in the field, the barriers start going up early on: many parents and schoolteachers are reluctant to press disabled children into subjects perceived to be challenging, such as science and mathematics. This is particularly unfortunate because these individuals may well have characteristics, such as problem-solving skills and perseverance, that would stand them in good stead in science.

Later on, the incentives and the equipment that could help individuals to function in the world of science are often lacking. In funding agencies and at universities, active support for the disabled as a group is sometimes patchy.

In the United States, the number of people with disabilities in higher education has tripled in the past 25 years, but they are still under-represented in science, technology, engineering and mathematics. According to the National Science Foundation (NSF), 7% of people in these areas have some degree of disability, compared with 13% of the working population as a whole. Figures for the severely disabled are hard to find, but anecdotal evidence suggests that this group is badly under-represented in science.

Some noteworthy efforts exist to recruit, train and assist students

with disabilities. The NSF's Research in Disabilities Education programme, for example, is spending \$3 million over five years to establishing 'regional alliances' to address the issue at a local level. The American Association for the Advancement of Science's Entry Point! programme collaborates with industry and government agencies to identify talented students and place them in paid summer internships. And NASA has recently established an initiative to help visually impaired students who would like to work there.

US institutions are legally obliged to make provision for the disabled under the landmark 1990 Americans with Disabilities Act. The European Union adopted a vaguer commitment to equal opportunities in 1996, but its application varies between nations. Mentoring of disabled scientists in Germany is encouraged, for example, by the Tandem Project at the Paul Ehrlich Institute in Frankfurt. In Britain, TechDis in York has prepared a host of valuable information for students with disabilities and their teachers. In addition, many colleges now include disability as a diversity issue, opening up funding opportunities previously allocated to women and ethnic minorities to people with disabilities.

Although laudable, these efforts are relatively small. Their limited scope contrasts with the extensive push that has been made to improve the opportunities for women and ethnic minorities in science. This is partly because the disabled are a smaller group, although not as small as you may think: there are an estimated 365,000 people with disabilities in science and engineering careers in the United States alone. But the more telling factor is the lack of direct political pressure on funding agencies to accommodate people with disabilities. That's a weak excuse for inaction in a sphere that cries out for some imaginative attention. ■

RESEARCH HIGHLIGHTS

Trick of the light

Preprint astro-ph/0505484 at <http://arxiv.org> (2005)
When instruments intended to detect light from planets outside our Solar System succeed in spotting distant pinpricks of light, the data should be interpreted with care.

Michael Jura of the University of California, Los Angeles, warns that some sources may be comets, not planets. He calculates that a comet like Hale-Bopp (pictured) passing through the planetary system of another star can reflect as much light as the planet Earth does, because of the vast amount of dust it sheds.

Tracking the course of possible extrasolar planets would allow true planets to be distinguished from comets, he says. Such observations might also illuminate the poorly understood nature of comet populations around other stars. Both NASA and the European Space Agency have plans to use space-based observatories to study extrasolar planets.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

R. RESSMEYER/GETTY IMAGES

GENETICS

Barely there

Science doi:10.1126/science.1113485 (2005)

The genome of an ancient cave bear has been partially reconstructed by a team headed by Edward Rubin of the Lawrence Berkeley National Laboratory. The study shows that degraded DNA can be pieced together from fossils, raising hopes of recovering the genomes of early hominids.

The team analysed all the genetic material in samples from a 40,000-year-old bear tooth and bone. Some 90% of the DNA analysed in this metagenomic approach was from bacteria or unidentified sources. But similarities between the bear genome and that of the modern dog meant the team could identify 26,861 base pairs of cave bear sequence.

ANIMAL BEHAVIOUR

Fruitless frolics

Cell 121, 785–794 (2005)

Cell 121, 795–807 (2005)

A single gene found in the neurons of fruitflies, called *fruitless*, controls sexual behaviour so strongly that female flies engineered to manufacture male *fruitless* proteins behave like male flies.

Normally, female flies do not initiate reproduction. But the engineered females target other female flies with a complex mating ritual. Equally, male *Drosophila melanogaster* engineered to lack these proteins become passive in courtship. The researchers who conducted these

experiments, led by Barry Dickson of the Austrian Academy of Sciences in Vienna, also discovered *fruitless* to be expressed in olfactory–sensory neurons in males.

NEUROBIOLOGY

Within sight

Neuron doi:10.1016/j.neuron.2005.04.023 (2005)

What we notice in the corner of our eyes can be determined by unnoticed features of objects in our focus, say Zoltán Vidnyánszky of Semmelweis University in Budapest and his colleagues. They got subjects to focus on coloured dots on a screen that moved too subtly for their motion to be detected. Dots were then introduced in the subjects' peripheral view, and people were best at picking these up when they moved in the same direction as the dots in plain focus. This suggests that objects on the edge of the visual field catch our eye when they share a common trait with those in our direct gaze.

L. MONNERET/GETTY IMAGES

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

CHEMISTRY

Maximum mass

Anal. Chem. doi:10.1021/ac0482054 (2005)

Biological molecules weighing a million times more than a hydrogen atom can now be detected using mass spectrometry, report Renato Zenobi and colleagues from the Swiss Federal Institute of Technology in Zurich. The technique — using a modified version of a MALDI-TOF spectrometer — could be applied to clinical analyses of viral and bacterial particles.

EVOLUTION

Love bug

Biol. Lett. doi:10.1098/rsbl.2005.0296 (2005)

It has been suggested that parasites are a key driver in the evolution of sexual reproduction. Support for this idea now comes from experiments in flour beetles (*Tribolium castaneum*) led by Paul Schmid-Hempel of the Swiss Federal Institute of Technology in Zurich. The beetles are locked in an evolutionary battle with the parasite *Nosema whitei*.

The researchers looked at two sequences in the beetle's genome, and they report that infection with co-evolving parasites increased shuffling of genes at these sites during meiotic cell division (by which eggs and sperm are produced). Such genetic recombination can break up favourable genetic sequences, but it may also give the beetle an advantage over the ever-changing parasite.

RNA INTERFERENCE

Meeting places

Nature Cell Biol. doi:10.1038/ncb1265 (2005)

Nature Cell Biol. doi:10.1038/ncb1274 (2005)

Genes are transcribed into messenger RNAs that are then translated into proteins. This central dogma of molecular biology has recently been shaken by the discovery of new classes of small RNAs. Micro RNAs and small interfering RNAs, for example, suppress gene expression without coding for proteins.

Two independent teams, George Sen and Helen Blau of Stanford University, and a group led by Gregory Hannon of Cold Spring Harbor Laboratory, have shed more light on how, or at least where, this happens. Both groups show that components of the cellular machinery needed to help these small RNAs repress the translation of mRNAs, concentrate in the P-bodies of mammalian cells — tiny organelles in the cytoplasm where superfluous mRNAs are processed. Whether this co-localization is a cause or a consequence of mRNA suppression has yet to be determined.

COMPUTATIONAL BIOLOGY

Patterns pending

J. Biol. doi:10.1186/jbiol23 (2005)

To help them build computer models of networks, such as the interactions between molecules inside cells, biologists developed the idea of 'network motifs'. Motifs are simple patterns of interconnection that occur more often than would be expected by chance.

Researchers led by Frederick Roth at Harvard Medical School in Boston investigated five networks that connect genes and proteins in cells of the yeast *Saccharomyces cerevisiae*. As well as finding motifs, they put motifs into higher-order groupings called 'network themes', which seem to represent the elemental design principles of the network. Maps based on such themes could predict the function of unknown genes.

PHYSICS

Speed trap

Phys. Rev. A **71**, 050101 (2005)

Is the speed of light the same in different directions? Yes, according to the most definitive answer yet given to this long-studied question.

Stephan Schiller's group at the Heinrich Heine University in Düsseldorf, Germany, searched for changes in the speed of light by studying laser beams that were reflected back and forth in sapphire crystals held close to

absolute zero. As the crystals were rotated, their resonance frequencies remained constant, suggesting that the speed of light is unchanged. The experiment's accuracy of around 6 parts in 10^{16} is an order of magnitude better than the most stringent previous test.

PALAEOLOGY

Tyrannosaurus sex

Science **308**, 1456–1459 (2005)

Examination of a *Tyrannosaurus rex* unearthed in 2003 in Montana has revealed a bone tissue similar to that of today's female birds, say researchers led by Mary Schweitzer from North Carolina State University, Raleigh.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Medullary tissue is found on the inside of hollow bones and is pitted with channels through which blood vessels run. This makes it an excellent source of the calcium that female birds need to produce eggshells. Evidence that *T. rex* used a similar system emphasizes how closely birds are related to these dinosaurs, and suggests that this specimen was an egg-laying mother.

MATERIALS

Going soft

Phys. Rev. Lett. **94**, 205502 (2005)

You can squeeze and mould it like putty. But the material developed by Wei Hua Wang, of the Institute of Physics in Beijing, and his team is not a plastic made from polymers — it is a metallic glass that goes soft in boiling water.

Wang's blend of cerium, aluminium and copper, with a dash of niobium, is completely amorphous. At room temperature, it has the hardness, toughness and electrical conductivity of a typical metal. When heated, it displays plastic properties, becoming malleable at just 68 °C. This unusually low glass transition temperature, combined with resistance to crystallization, means the material should appeal to manufacturers.

JOURNAL CLUB

Luigi Piro
National Institute for
Astrophysics, Rome

The mission scientist for a space telescope that probed the origin of gamma-ray bursts explains why these dramatic pulses of energy have cosmic importance.

I was the lead scientist for BeppoSAX, a satellite that could pinpoint gamma-ray (γ -ray) bursts. We discovered that these energy bursts are produced when massive stars in the distant Universe undergo powerful explosions. Intriguing questions remain about the nature of such events, but the bursts also provide thrilling data for cosmologists.

My team and other astronomers are considering how to use γ -ray bursts to learn about the early history of the Universe. One idea is to analyse the accompanying X-rays to get information about the composition of the regions through which the radiation has travelled. Another idea is to use γ -ray bursts as 'cosmic rulers': if astronomers know a burst's total energy, then its brightness, as viewed from Earth, can be used to infer its distance.

Cosmology was revolutionized when type 1a supernovae — exploding stars — were used as cosmic rulers. They revealed that the current expansion of the Universe is accelerating. Gamma-ray bursts should tell us more because they are brighter, and so reach us from farther away.

Indeed, a report in *The Astrophysical Journal* (G. Ghirlanda, G. Ghisellini, D. Lazzati and C. Firmani **613**, L13–L16; 2004) shows that using even a limited number of γ -ray bursts as rulers reduces the margin of error in current cosmological models. The researchers calculated the bursts' total energies from their peak frequencies and directions.

Observations of γ -ray bursts from Swift, a new space telescope, will further test this method and, I hope, shed light on dark energy. Dark energy is thought to drive the expansion of the Universe, but its physical nature is a mystery.

PHOTOLIBRARY.COM

NEWS

Chemistry society goes head to head with NIH in fight over public database

WASHINGTON DC

Many chemists might not know it, but the organization that represents them in the United States is fighting to limit their free access to chemical information. The American Chemical Society says that a new publicly funded database of molecules threatens its own fee-based Chemical Abstracts Service (CAS), and it is lobbying politicians to restrict the free version. But it is having trouble convincing members that this is in their interests.

CAS is part of many chemists' daily routine. The service is a massive registry of chemicals with their structures and properties, as well as links to related publications and patents.

Depending on their size, chemistry departments and companies pay from a few thousand to more than a hundred thousand dollars for a year's access to the database.

Chemists have had no alternative. A journal search will not find a chemical structure, so the database is the only way to find previously reported molecules and reactions, short of wading through papers by hand.

"CAS is very important," says Chris Reed, an inorganic chemist at the University of California, Riverside. "My students use it all the time, for mining the literature or finding the compounds they want."

PubChem, a free database launched by the US National Institutes of Health (NIH) last September, threatens CAS's monopoly. It is smaller, containing 650,000 molecules so far compared with CAS's 25 million. And it is aimed more at biologists, linking to information such as gene sequences, and related papers in the NIH's PubMed archive of biomedical journals.

650,000 and rising

But it is growing. On 25 May, records were added from NMRShiftDB, a database of chemicals' nuclear magnetic resonance spectra, and from *Nature Chemical Biology*, which requires all authors to submit their data to PubChem. Other sources are likely to follow.

The ACS argues that projects that compete with the private sector are a waste of taxpayers' money. The database generates the lion's share of the non-profit ACS's income of \$375 million, which pays for the society's publications, meetings and staff.

So the society is trying to persuade Congress to make the NIH restrict its database to molecules found by NIH researchers.

Steve Bryant, project director for PubChem, says that's unfair, because the linked content provided by the two databases is different, and they serve different audiences.

Bob Massie, head of CAS, disagrees. "We have been hearing that every chemical researcher understands that PubChem is a substitute for CAS," he says.

To try to limit PubChem to information produced by NIH researchers, the ACS has been working with lawmakers in Ohio, where CAS employs almost 1,300 people. In particular, it has lobbied congressman Ralph Regula (Republican, Ohio), the chairman of the appropriations subcommittee that allocates money to the NIH.

The society's efforts have intensified ahead of this week's expected debut of the 2006 House Appropriations bill that outlines the agency's proposed budget. As *Nature* went to press, the draft bill was due on 9 June. An official report accompanying the bill was expected to ask the NIH to limit PubChem to data produced by its own efforts. The report is not legally binding, but if the bill is passed it would be difficult for the NIH to ignore.

Although many chemists are unaware of the ACS's attempt to restrict PubChem, weblogs and library discussion groups have picked up the subject. The fight is turning sour. "My only interpretation of the ACS's recent actions is that it is no longer trying to represent the best interests of the scientists who form its

One in three scientists confesses to having sinned

More than a third of US scientists, in a survey of thousands, have admitted to misbehaving in the past three years. The social scientists who carried out the study of research misconduct warn that because attention is focused on high-profile, serious cases, a broader threat from more minor deeds is being missed.

Their conclusions may hit a nerve, particularly among scientific societies in the United States. Throughout the 1990s, these groups fought to limit their government's definition of misconduct and the types of behaviour it is responsible for policing.

Brian Martinson of the HealthPartners Research Foundation in Minneapolis, Minnesota, and his colleagues mailed an anonymous survey to thousands of

scientists funded by the National Institutes of Health. They asked the scientists whether they were guilty of misbehaviours ranging from falsifying data to inadequate record keeping.

Of 3,247 early- and mid-career researchers who responded, less than 1.5% admitted to falsification or plagiarism, the most serious types of misconduct listed. But 15.5% said they had changed the design, methodology or results of a study in response to pressure from a funding source; 12.5% admitted overlooking others' use of flawed data; and 7.6% said they had

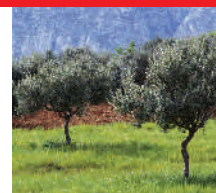
circumvented minor aspects of requirements regarding the use of human subjects (see page 737).

Overall, about a third admitted to at least one of the ten most serious offences on

"The majority of misbehaviours reported are more corrosive than explosive. That makes them no less damaging."

the list — a range of misbehaviours described by the authors as "striking in its breadth and prevalence".

But Arthur Caplan, director of the Center for Bioethics at the University of Pennsylvania, Philadelphia, cautions against concluding that the structure of science is corroded. He points out that dropping an outlying data point is not the same as plagiarizing a paper.



PLANT CARBS POWER CARS

Fuel born from carbohydrates could be clean and easy.

www.nature.com/news

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Pricing structure: the American Chemical Society makes much of its money selling molecular data.

membership," says Richard Roberts, a chemist at New England Biolabs in Beverly, Massachusetts, and 1993 Nobel laureate, who advises PubChem. "Rather, it seems to be a commercial enterprise whose principal objective is to accumulate money." To protest, he has pulled out of an upcoming ACS conference in India.

Roberts and the NIH are spreading misinformation, responds Madeleine Jacobs, executive director of the ACS. She complains that NIH staffers have implied that the ACS is trying to shut PubChem down. "We do not and never have opposed the concept of publishing data generated by NIH grantees," she says.

"I don't mean to say that the problems identified don't merit deliberation and a response," he says. "But there may be a tendency if you just read the headlines to say, 'Oh my goodness, the ethical house of science is collapsing around us'."

Martinson counters that, although individual cases may not be as serious as fraud, the survey reveals a threat to the integrity of science that is not captured by narrow definitions of misconduct. "The majority of misbehaviours reported to us are more corrosive than explosive," he says. "That makes them no less damaging."

He thinks the main cause of all the questionable behaviour is the increasing pressure that scientists are under as they compete to publish papers and win grants.

"We need to think about the working conditions in science that can be addressed," he says, suggesting better salaries and employment conditions for young scientists, and a more transparent peer-review process.

He is at pains to stress that he does not think governments should expand regulation of scientific behaviour. And when it was shown Martinson's study, the Federation of American Societies for Experimental Biology, based in Bethesda, Maryland, was quick to reiterate its support for the narrow definition of misconduct that was officially agreed in 2000.

"The US government adopted 'fabrication, falsification and plagiarism' as the defining criteria, a policy with which we

"But what we are seeing now goes far beyond what the NIH first proposed."

Jacobs points out that most of the money raised from CAS goes towards the society's publishing services, and that these would be jeopardized if profits fell. ACS members she has spoken to agree that PubChem is a waste of public money, she adds.

Offshore option

Not everyone agrees. "Most of the members I've spoken to are kind of upset about it," says Reed. He is drafting a letter to *Chemical & Engineering News*, an ACS publication, to complain about the society's actions. Although he doesn't use PubChem himself, he objects to any attempt to squash it. "I can understand the society being nervous about competition, but I think something that is complementary and even a bit competitive is healthy."

He says that he and his colleagues would like to see the ACS be "more forward and innovative in opening access to databases and literature". The University of California is disseminating information on the quarrel to its chemists, along with suggestions for action, and Regula's phone number.

Meanwhile, a group of European chemists, including Peter Murray Rust of the University of Cambridge, UK, is taking a different approach. Worried that researchers elsewhere would lose out if information is removed from the NIH site, they are discussing setting up a European-funded mirror of the site with PubChem, which the US government would have no power to restrict.

Emma Marris

concur," says Paul Kincade, the federation's president. That means the government cannot investigate or punish any behaviours outside that definition.

In 2002, scientific societies led by the federation and the Washington-based Association of American Medical Colleges fought a government office's plan to collect data on such behaviours (see *Nature* 420, 739–740; 2002). The societies argued such monitoring should be the responsibility of scientists themselves.

Martinson and his colleagues say their study is the first attempt to quantify such activities. They hope their results will persuade scientists to stop ignoring the wider range of misbehaviour.

Meredith Wadman

Blue Brain boots up to mixed response

One of the boldest brain-modelling projects ever attempted is about to get under way in Switzerland. A team of neuroscientists plans to use a supercomputer to create a biologically realistic simulation of the neural circuits responsible for higher mental processes in humans and other mammals. But experts elsewhere are divided about its chances of success.

Most brain models focus either on large- or small-scale features. Some teams have connected huge networks of simple units to recreate brain functions such as vision, for example, whereas others have built detailed computer simulations of individual neurons. Now a team from the Swiss Federal Institute of Technology in Lausanne (EPFL) will use the IBM supercomputer Blue Gene/L in an attempt to combine both approaches.

"This is one of the most ambitious computational-neuroscience projects ever planned," says Alain Destexhe, a modeller at CNRS, France's main basic-research agency, in Gif-sur-Yvette. "Blue Gene is one of the most powerful computers ever made available to neuroscience."

The Blue Brain Project will simulate part of the neocortex, the intricately folded region on the outside of mammals' brains. The Swiss team will model a column of cells from the rat neocortex, the animal for which the most detailed data are available. Each column is just 2 millimetres high and half a millimetre in diameter, but contains some 10,000 cells connected by 5 kilometres of fibres. In both humans and rats, these columns form a basic circuitry that is repeated across the cortex and controls everything from vision and movement to planning.

"It's as if in evolution this system was cloned

and duplicated in mammals — it makes up 80% of the human brain," says Henry Markram, who leads the project at the EPFL. "The more we look at the neocortex, the more we are in awe of it."

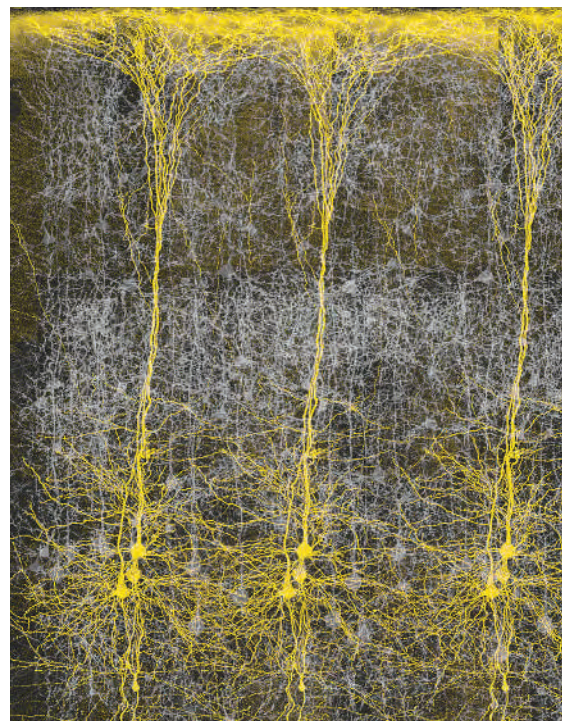
IBM has built several Blue Gene computers for use in research (see 'Virtual Big Bangs and digital mushroom clouds'). The EPFL's version, which is worth US\$10 million and should be switched on next month, is a rack of four refrigerator-sized units, each containing more than 1,000 processors. It can process over 22 trillion operations per second. If running today, it would probably be the fourth most powerful machine in the world.

Markram will feed the simulation with information from his group's database on the

"Half the research community will say this is nonsense."

physiological and electrical properties and shapes of cortical cells, which is regarded as the most complete database of this kind in the world. Each Blue Gene processor will model one or two individual cells, which will then be connected up as in the rat cortex. The team will spend about three years testing and refining the system by comparing the results of simulations with experiments using real tissue.

Such a grand project is bound to attract critics, acknowledges Markram. "Half the research community will say this is nonsense," he says. Sure enough, when *Nature* put the idea to brain modellers, they hailed Blue Brain as a great leap forward in terms of the realism of simulations, but some doubted whether



Makes you think: neuroscientists aim to create a computer model of a column of cortical cells.

BRAIN MIND INSTITUTE

Markram has enough data on the cortex to make his plan work.

"This is an ambitious project that is bound to fail," says Terry Sejnowski of the Salk Institute for Biological Studies in San Diego, California. "We are still far from understanding enough about the brain to build a detailed realistic model."

Neuroscientists say that too little is known

Europe halts decisions on stem-cell patents

MUNICH

The European Patent Office (EPO) has put patent applications involving human embryonic stem-cell technology on ice. And there are no immediate prospects for a thaw.

The EPO president, Alain Pompidou, said last week during the presentation of the organization's 2004 report that the EPO will not patent any embryonic stem-cell technology for the time being, because "there are too many ethical aspects that have not been resolved at the political level".

The EPO and the European Union have different members and are ruled by different conventions. But the EPO needs to take note of the European Union's political climate, Pompidou said. The European Commission is deciding whether to fund research on human embryonic stem cells.

Because of the moratorium, the EPO has so far received only three applications involving human embryonic cells; none has yet been approved. The office has been inundated with appeals against them, and

officials are concerned that any patent granted would trigger protests similar to those that surrounded patents for genetically modified organisms. In February 2000, activists from the environmental group Greenpeace bricked up the door to the organization's Munich headquarters.

The EPO worries that it would not get political backing to defend itself against such action in the wake of a human embryonic stem-cell patent.

Many European scientists are unhappy

Virtual Big Bangs and digital mushroom clouds

Blue Gene/Ls are bringing huge computing power to fields from cosmology to drug discovery, because they are relatively cheap and compact.

WEAPONS RESEARCH

Lawrence Livermore National Laboratory in California has the cream of the Blue Gene crop: a 64-rack machine that IBM says will be the fastest computer in the world. Due for completion this summer, the device will be used for nuclear-weapons research.

RADIOASTRONOMY

The Netherlands' Low Frequency Array radio telescope will use a four-rack Blue Gene machine based at

the University of Groningen. It will analyse data from 15,000 antennas spanning an area 350 km in diameter.

PROTEIN FOLDING

Researchers at the University of Edinburgh, UK, are using a single-rack Blue Gene machine, capable of 6 trillion operations per second, to simulate protein folding and fluid mixing.

CLIMATE RESEARCH

A single-rack Blue Gene is being shared by the National Center for Atmospheric Research and the University of Colorado, which will run simulations of ocean, weather and climate behaviour.

COSMOLOGY

A team at the San Diego Supercomputing Center is using a single-rack Blue Gene to run Enzo — the centre's software that simulates how galaxies evolved from the Big Bang. The simulation, which has already predicted how the first stars formed, produced around 30,000 gigabytes of data on a typical recent run.

DRUG DEVELOPMENT

Japan's National Institute of Advanced Industrial Science and Technology is using its four-rack device to boost drug development. Researchers there hope to shed light on how drugs interact with their targets in the body.

about the structure of the network connecting cortical cells, for example. They add that a truly realistic model would have to incorporate molecular activity in the regions where neurons connect, a level of detail that is currently beyond the Blue Brain Project.

Markram agrees that more data are needed, especially from the bottom section of the six-layer cortical column, but says that work to remedy this will be done while the machine is being tested. Once the model is fully functional, researchers will be able to play with it, tweaking the numbers of different types of neuron, for example, or altering the levels of certain neurotransmitters, to see how such changes affect larger-scale brain activity.

The machine could also probe conditions in which cortical circuitry seems to malfunction.

For example, some symptoms of autism can be recreated in rodents by giving thalidomide during pregnancy. The thalidomide causes changes in the cortex cells that could be mimicked on Blue Brain, says Markram.

In the long-term, Markram has a grander plan that will raise eyebrows among even his most supportive peers. Within ten years, he predicts, column models could be duplicated and connected to create simulations of the whole cortex and eventually the whole brain.

Fred Wolf, a computational neuroscientist at the Max Planck Institute for Dynamics and Self-Organisation in Göttingen, Germany, is enthusiastic about the idea. But he adds: "Don't expect to see whole-brain simulations any time soon."

Jim Giles

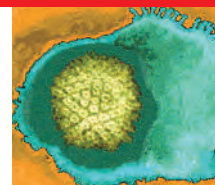
with the delay. Patent offices in the United States and many Asian countries, including South Korea, allow inventions involving human embryonic stem cells to be patented. "The EPO underestimates the economic damage this attitude will cause in Europe," warns Oliver Brüstle of Germany's Institute for Reconstructive Neurobiology, one of the three applicants.

One hope remains for Brüstle and his colleagues. One of the applications, rejected last year, is making its way through the

EPO's appeals procedure. It is being considered by the technical appeals board, which will give its judgement in October.

Siobhán Yeats, director for biotechnology at the EPO, says that if the board cannot settle the case on technical grounds, it may pass it on to the EPO's highest appeals board, the Enlarged Board of Appeal. The Enlarged Board rules on the fundamental patentability of inventions, and so could set a precedent.

Sonja Schubert



PAINKILLER
Deactivated virus
boosts relief for
sensitive nerves
www.nature.com/news

SPL

Soap opera reaps prize for its clean message

WASHINGTON DC

A Vietnamese soap opera is one of the winners of a World Bank competition for ideas to improve the livelihoods of people in the developing world.

The radio drama aims to teach Vietnamese rice farmers about sustainable farming, by weaving pest-management principles into the loves-and-hates storyline of a typical soap opera (see *Nature* **430**, 284; 2004).

The response to the soap opera, which is based on the UK radio drama *The Archers*, has been overwhelming, says Monina Escalada, a research fellow at the International Rice Research Institute in Manila, the Philippines. Escalada organizes the writers and environmental scientists who team up to produce the scripts.

"We know farmers are listening," she adds. With the prize money, she says, the drama will be expanded to include other environmental issues, such as reducing straw-burning and conserving water.

The drama was one of more than 2,600 proposals submitted to the World Bank's Development Marketplace grant programme. The 78 finalists, which included individuals,

governments, private foundations and academics from 42 countries, were invited to Washington DC last month to present their projects.

The theme of this year's competition was the environment, and 31 winning projects were chosen. They ranged from a Kenyan plan to pay the poor to collect charcoal that can be turned into clean-burning briquettes, to using solar-powered ovens in Costa Rica to sterilize hospital waste, which is normally dumped in landfills without treatment. The winners received up to \$150,000 each, with the remaining finalists each getting \$5,000.

The Development Marketplace started in 1998 as an internal competition "to get innovative, bottom-up ideas from staffers", explains John Wilton, vice-president of the bank's strategy, risk and finance group. But in 2000 it was opened to international entrants. "We soon realized that it would be more exciting and useful to have the competition come from our client countries."

Since then, nearly US\$35 million has funded more than 570 projects in some 70 countries. Not all have gone on to do great things. But one of the big winners in the 2000 global contest was a South African proposal to harvest children's

energy by installing water pumps masquerading as merry-go-rounds in village schools.

As the children play, water is pumped into a storage tank. HIV/AIDS awareness messages line two sides of the tank and the remaining space is sold for advertisements, which pay for the pumps' upkeep. There are now more than 400 play pumps in South Africa, and the idea is spreading throughout the continent as well as to south Asia and Latin America.

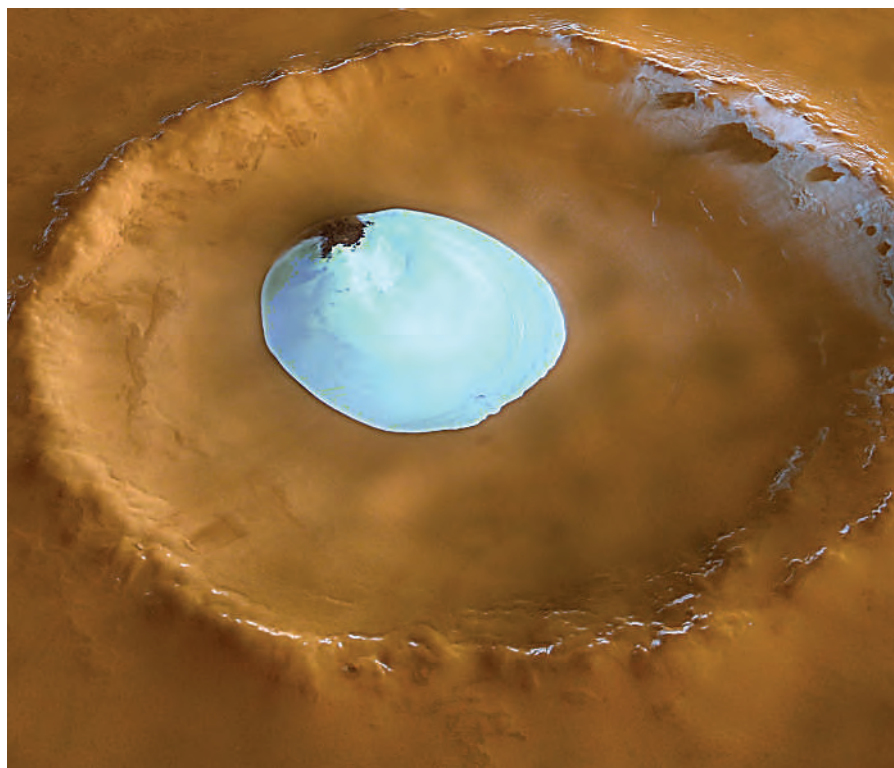
Observers welcome the competition's grassroots approach. "This is an excellent initiative to support local, innovative projects," says Manish Bapna, head of the Bank Information Centre, a watchdog of the World Bank.

But Bapna and others are also quick to point out that the investment involved is tiny compared with the money that they believe the World Bank spends on projects that are in the interests of big business, rather than local people or the environment.

"This reveals once again the inherent inconsistencies in World Bank activities," Bapna says. "It remains a lead financier for environmentally controversial oil, gas and mining projects."

Jessica Ebert

ESA/DLR/FREE UNIV. BERLIN (G. NEUKUM)



SNAPSHOT

Sunlight on an icy martian crater

This image from the Mars Express spacecraft shows a pocket of water ice nestling in a martian crater, bathed in the late martian summer sun.

The shadow of the crater's rim, which towers 300 metres over the surrounding plains, prevents the ice from vaporizing in the planet's thin atmosphere. A dusting of frost survives inside the rim to the upper right, while the sun glimmers on its south-facing outer edge.

The 35-kilometre-wide crater sits 70° north of the martian equator, in a low-lying region known as Vastitas Borealis. Previous orbiters have spotted ice deposits in craters, but the High Resolution Stereo Camera on board the European probe is the first to return a three-dimensional colour image of an icy spot. The ice may be up to 200 metres thick, and lies over a dune field that has formed in the sediment on the crater's floor. The data were collected on 2 February, and this image was created for *Nature* last week.

ON THE RECORD

“We tend to eat a lot of fried foods...most restaurants don't have healthy choices.”

West Virginian Rudy Philips tells *The New York Times* why the US Centers for Disease Control and Prevention has launched its first-ever investigation into an ‘outbreak of obesity’ in his state.

“I'm sure skinny people go to those same restaurants.”

Daniel McGee, a statistician at Florida State University, finds Philips's explanation unpalatable.

SCORECARD**No jacket required**

Japan is rolling up its sleeves to hit its emissions targets. The government is asking businessmen to throw off their sharp suits and go for that cool, casual look to cut the use of air conditioning.

**Culture club**

Egged on by recent cloning success, South Korea is planning a world stem-cell bank. And with current rates of interest, the growth potential is enormous.

**Dope show**

Citing a possible link between homegrown cannabis and the global drugs market, the US Supreme Court has made medical marijuana illegal, prompting some to ask whether it has truly lost the pot.

OVERHYPED**Life on Mars**

When methane was found in Mars's atmosphere last year, the media (and many scientists) seized on the idea that it was a whiff of life from martian bacteria. As estimates of the gas grew, so did column inches on the hopes of finding microbes.

But the real source may be far more mundane: chemistry. Geologists have calculated that given a supply of water and carbonates, just 80,000 tonnes of the mineral olivine would be enough to generate a year's worth of methane. Another paper reports a Cuba-sized olivine field on Mars's surface, and suggests there may be more beneath. Could the prospects for life be stone dead?

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Thorny problem: Arizona's saguaro cacti can take decades to grow back after a desert fire.

Pall hangs over desert's future as alien weeds fuel wildfires

TUCSON, ARIZONA

The wildfire season in Arizona has begun early this year. Following unusually high temperatures around the end of May, smoke has been seen rising at many places in the Sonoran Desert, and fire warnings have been issued across the region.

Ecologists warn that the problem is about to get far worse. Unless urgent action is taken, they fear that the uncontrolled spread of exotic weeds could trigger an ecological disaster in the area, as the frequency of fires rises dramatically.

Until the arrival of alien weeds, wildfires were not a problem in the region, which stretches from the southwestern United States to northern Mexico. The native vegetation is not dense enough to burn well. But thickly growing non-native grasses, which dry up in summer, cover the ground in a layer that spreads fire easily. The weeds have also become a serious threat to desert biodiversity, as they compete with native plants such as the saguaro cactus, which can take decades to grow back after a fire.

A shrubby grass known as buffelgrass (*Pennisetum ciliare*) has become common in the desert since the 1960s, and is largely responsible for the region's increasing wildfire problem since then. But ecologists fear that another invading weed, Sahara mustard (*Brassica tournefortii*), could have an even more devastating effect.

“Sahara mustard is the greatest threat to the

Sonoran,” says Mark Dimmit, a botanist and director of natural history at the Arizona-Sonoran Desert Museum near Tucson. “If nothing is done, it could turn the desert into a wasteland.”

Earlier this year, the Arizona state government put buffelgrass on a list of prohibited species, which means planting and transporting the noxious plants and their seeds is banned. And in the Tucson area, where a wildfire threatened suburban neighbourhoods in 2003, attempts are under way to eradicate buffelgrass mechanically. The weed grows near roads

and highways, and in patches on desert hills and slopes.

But control is more difficult with Sahara mustard, which spreads extensively across valley floors, roadsides, rocky hillsides and even sand dunes.

Its deep roots are hard to remove mechanically, and there is no biological agent for its control. “It's everywhere now,” says Dimmit. “We really don't know what to do.”

To gauge the scale of the problem, the Arizona-Sonoran Desert Museum began training a team of ‘citizen scientists’ in April. These volunteers will be equipped with data sheets and digital cameras and are about to begin mapping the weeds' spread.

But as a desert ecologist, Dimmit wishes he could spend his time on other things. “Looking at weeds gets really boring,” he says. “Conserving good habitats would be much more exciting.”

Quirin Schiermeier

“If nothing is done, Sahara mustard could turn the desert into a wasteland.”



MARTIAN METHANE
Chunks of the mineral olivine are enough to explain 'whiff of life', claim geologists.
www.nature.com/news

NASA/JPL/ASU

Evolutionist row makes museum ditch donation

WASHINGTON DC

Is one of Washington's most prestigious museums promoting intelligent design by screening a film that some scientists claim is anti-evolution? The Smithsonian Institution's National Museum of Natural History has returned a \$16,000 donation it had accepted from a think-tank that promotes the philosophy, but it still plans to allow the group to screen its controversial movie.

The trouble began in April, when Smithsonian officials rented an auditorium to the Seattle-based Discovery Institute. The institute is a major centre of the intelligent-design movement, which holds that an intelligent creator, and not natural selection, shaped life on Earth.

Evolution advocates were shocked to receive invitations from the Discovery Institute that claimed the 23 June event was co-sponsored by the museum's director. "It looked as though the Smithsonian was

supporting intelligent design," says Eugenie Scott, head of the National Center for Science Education in Oakland, California, which works to improve the teaching of evolution.

According to spokesman Randall Kremer, the museum regularly accepts donations for use of the auditorium, and staff were unaware of the institute's philosophy.

"It was treated as a routine request," he says.

"The Smithsonian was duped," adds Lawrence Krauss, a theoretical physicist at Case Western Reserve University in Cleveland, Ohio, who lectures on evolution.

Discovery Institute staff deny any wrongdoing. "We actually followed the invitation template that the Smithsonian provided for us," says Jay Richards, a senior fellow with the institute.

Nevertheless, after dozens of calls and e-mails from researchers and the public, the museum decided last week to return the donation and issue a statement disavowing the event. "We are not in any way changing

the foundation of research here," says Kremer. However, the museum will honour its contract and allow the Discovery Institute to show *The Privileged Planet: The Search for Purpose in the Universe*.

Richards adds that the film does not deal directly with intelligent design or biological evolution. It is based on a book he co-authored with Guillermo Gonzalez, an astronomer at Iowa State University, which argues that Earth is uniquely and improbably suited for the appearance of life.

Scott says she is pleased by the swift steps the museum has taken. But not everyone is satisfied. "I personally don't think they should show the film," says Peter Folger, director of outreach for the American Geophysical Union in Washington DC, adding that the screening could give the institute scientific credibility.

The museum is reviewing its special-events policy to avoid confusion in future. ■
Geoff Brumfiel

Germany casts doubt on extra Framework funding

Plans to double the European Union's research budget may have suffered a major setback last week when Germany indicated that it will back down from its proposed funding.

In April, the European Commission proposed a budget of €68 billion (US\$83 billion) for the Seventh Framework Programme, to run from 2007 to 2013. But at a meeting in Berlin on 2 June, Joschka Fischer, the German foreign minister, told Janez Potočnik, the European commissioner for research, that Germany cannot afford to finance any substantial increase in European budgets.

Europe's heads of governments will discuss European budgets at a summit meeting next week in Brussels, but the commission expects the initial proposal to be cut by more than a third. "Obviously, we are in a very, very difficult situation," said Peter Dröll, head of Potočnik's cabinet, at a workshop of research funding experts on 3 June.

Libya delays verdict on AIDS nurses' appeal

Five Bulgarian nurses and a Palestinian doctor sentenced to death in Libya in May 2004 will remain in jail for six more months before knowing the outcome of their appeal. A Libyan court found the group guilty of deliberately infecting hundreds of children with HIV, the virus that causes AIDS.

The six have languished in jail since 1999. They claim they are innocent and that their confessions were extracted under torture. Scientists have taken up their case, saying the evidence shows that many of the children were infected as early as 1994, long before the workers arrived in Libya (see *Nature* 430, 277; 2004). The decision by Libya's supreme court to delay the verdict from 31 May until 15 November came after a visit by Benita Ferrero-Waldner, the European Union's commissioner for external relations.

Treaty on transgenics hits trouble over labelling

A major conference meant to iron out differences over shipping genetically modified (GM) organisms between countries has ended in failure.

Negotiators left Montreal last week in disappointment after arguing over the Cartagena Protocol on Biosafety, which is meant to ensure the safe international handling of GM organisms. The stumbling



Japan plays trump card to get kids into science

Japanese biologists have a new way to get kids interested in science — a card game based on manga, the cartoons that Japan has exported to the world. The RIKEN Center for Developmental Biology in Kobe developed the game, which resembles 'card battle' games such as Pokemon. But in this game, players pick up information and images about developmental biology, such as *Drosophila* mutations and developmental stages.

The game gives scientists a manga-style makeover. From an enthusiastic postdoc to a lovely pipette-wielding technician, the characters

are colourful and attractive. "It would be great if we can help supplant some of the stereotypes about scientists before they take root," says Doug Sipp, manager of the centre's Office for Science Communications and International Affairs, "but really we just wanted to make something kids can collect and play with."

Students visiting the centre will get the cards for free. English versions will be distributed at some scientific meetings, including the Society for Developmental Biology meeting in San Francisco this July.

block was an article describing the documentation needed to ship them. Exporting countries prefer notification that a shipment "may" contain GM products; opponents demand detailed documents on the specific genetic alteration.

The group will now miss its September deadline, leaving countries with no international framework to refer to as they develop their own biosafety regulations.

Schwarzenegger breaks rank on climate change

California governor Arnold Schwarzenegger wants the state to slash its greenhouse-gas emissions, a move that challenges the Bush administration's rejection of the Kyoto Protocol on climate change.

Last week, after receiving a letter from 500 members of the Union of Concerned Scientists, Schwarzenegger announced that

California would aim to cut emissions to 20% of their 1990 levels by 2050. His proposal contains few specific measures, but is thought to include incentives for businesses to cap and trade emissions.

NASA gives go-ahead to Juno journey to Jupiter

Jupiter beat the Moon last week to become the destination for NASA's next New Frontiers mission. The Juno spacecraft, led by Scott Bolton of the Southwest Research Institute in Boulder, will be launched by 2010 into orbit around Jupiter to study the planet's interior and atmosphere.

Juno beat a proposal called Moonrise, which would have flown two landers to the lunar south pole and brought back rock samples. Juno had the edge in part because another unmanned NASA Moon programme is under way, whereas opportunities to visit the outer Solar System are rare.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Leading role: Arnold Schwarzenegger commits California to slashing greenhouse-gas emissions.

Correction

The News Feature "Consenting Adults? Not Necessarily..." (*Nature* 435, 138; 2005) quoted Xiaomei Zhai, a bioethicist at Peking Union Medical College, as saying: "For AIDS, we have to accept that we have to compromise." This quote was taken out of context. Zhai was not justifying a general compromise in ethical standards for clinical trials; rather, she was pointing out that the lack of suitable animal models makes it necessary to initiate research directly on human subjects. *Nature* apologizes for this error.

High-risk physics

In the world of billion-dollar particle physics, one lab is taking an enormous gamble on its future. **Geoff Brumfiel** takes a look at Fermilab's hopes to host the next big machine.

Sitting in a temporary office, Pier Oddone seems relaxed, even confident. Next month, the 60-year-old physicist takes the helm of the leading high-energy physics laboratory in the United States, at a time when the lab's particle accelerator — its *raison d'être* — is scheduled to close within five years. But Oddone is optimistic about the lab's future. "It is a tremendous opportunity to move vigorously in new directions," he says.

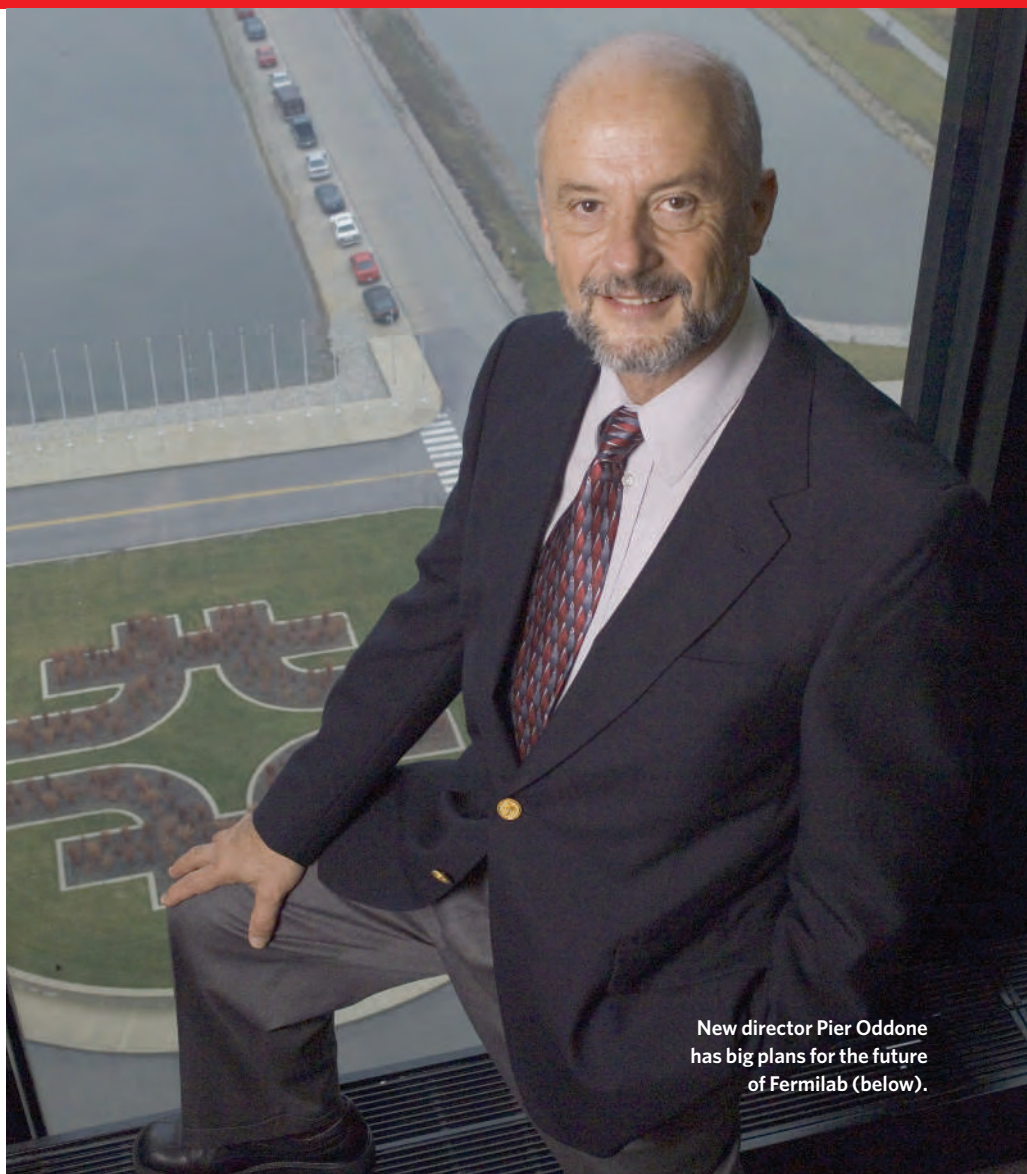
Oddone will be the fifth director of the Fermi National Accelerator Laboratory, or Fermilab, a 2,100-person government facility located in Batavia on the edge of the Illinois prairie. The lab is home to the Tevatron — the most powerful particle collider in the world today. But within three years or so it will be supplanted by a bigger machine at CERN, the European high-energy physics laboratory near Geneva. When that happens, the US government plans to switch off the Tevatron, leaving Fermilab with a very uncertain future.

Fermilab's full-time employees are understandably feeling nervous, says Robert Roser, a physicist and spokesperson for the Collider Detector Facility (CDF), one of the Tevatron's detector experiments. "If it doesn't have any physics going on, a lab like this could be very tempting to shut down," he says bluntly.

Making plans

Oddone's hope is to make Fermilab the home to a newer, bigger accelerator called the International Linear Collider. This would be the largest collider yet: 30 kilometres long and capable of smashing electrons together at energies of 500 billion electronvolts (eV) using state-of-the-art superconducting technology. It's so ambitious that physicists know it will take an international effort to build it. It would also cost a whopping US\$6 billion, according to more conservative estimates.

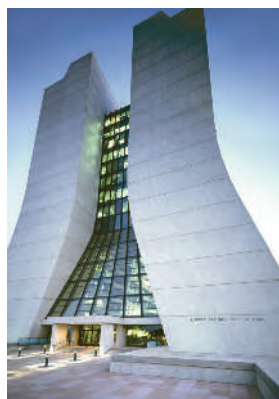
Oddone wants to work with international partners to begin building the linear collider at Fermilab by the end of the decade. To pull it off will require scientific cooperation, international diplomacy and a huge financial commitment from the US government.



New director Pier Oddone has big plans for the future of Fermilab (below).

The plan is a gamble, one that could restore Fermilab to its former glory, or leave it adrift without a clear mission or future. Critics say that laboratory administrators are staking too much on a machine that has yet to win political support in the United States. But supporters say that to win big you have to dream big.

Physicists started building the Tevatron in 1979 with just such big dreams in mind — among them the goal of confirming the existence of the top quark. Quarks are the building blocks that make up protons, neutrons and other subatomic particles and the top and bottom quarks are the heaviest. The bottom quark had been detected at Fermilab in 1977, using the existing main ring accelerator. To find the heavier top quark meant smashing protons and anti-protons together at energies between 500 million and 1.5 billion eV, which the Tevatron was designed to do.



With further modifications, the collider netted the top quark in 1995. This discovery created elation among physicists and a headache for Fermilab management: what next?

Fermilab was not the only US particle-physics lab facing this dilemma. Nobel laureate Burton Richter, director emeritus at the Stanford Linear Accelerator Centre (SLAC) in Menlo Park, California, says he had a similar problem in 1992. The lab's electron accelerator, which was the first to detect quarks in 1968, was at the end of a long line of upgrades, and there was no space to expand.

So, Richter says, they decided to convert the accelerator into a powerful X-ray source, which could be used by biologists, chemists and materials scientists to determine the structure of molecules and materials. SLAC also began to diversify its research, moving into astrophysics, γ -ray detection and cosmology. Today the lab is thriving and its budget is rising

steadily. Just last month SLAC announced a major restructuring to complete its transition to a multipurpose facility. "The new structure is adapted to allow our scientists to get on with what they do best — make major discoveries," says SLAC director Jonathan Dorfan.

Rather than diversifying, Fermilab took the decision in the 1990s to upgrade the Tevatron to even higher energies. This time their quarry was the Higgs boson, which, if theorists' predictions are correct, helps explain why everything in the Universe has mass. Bagging the Higgs would be a huge coup for Fermilab.

The US government spent hundreds of millions of dollars on the latest upgrade, which was completed in 2001. But the ageing Tevatron wasn't up to the task: its foundations had sagged over the years, and some of the new technology intended to boost its power behaved unpredictably. Budgets ran tight, and progress was slow. Today, after four difficult years, things are going smoothly, but so much time has been lost that few believe the Tevatron will discover the Higgs particle before CERN's machine is switched on in 2008.

Which leaves Fermilab in limbo, according to Richter. A year after CERN's machine turns on, the Tevatron will switch off, leaving lots of physicists with little to do, he says: "They have gotten themselves in a peculiar position, now the question is how do they get out of it?"

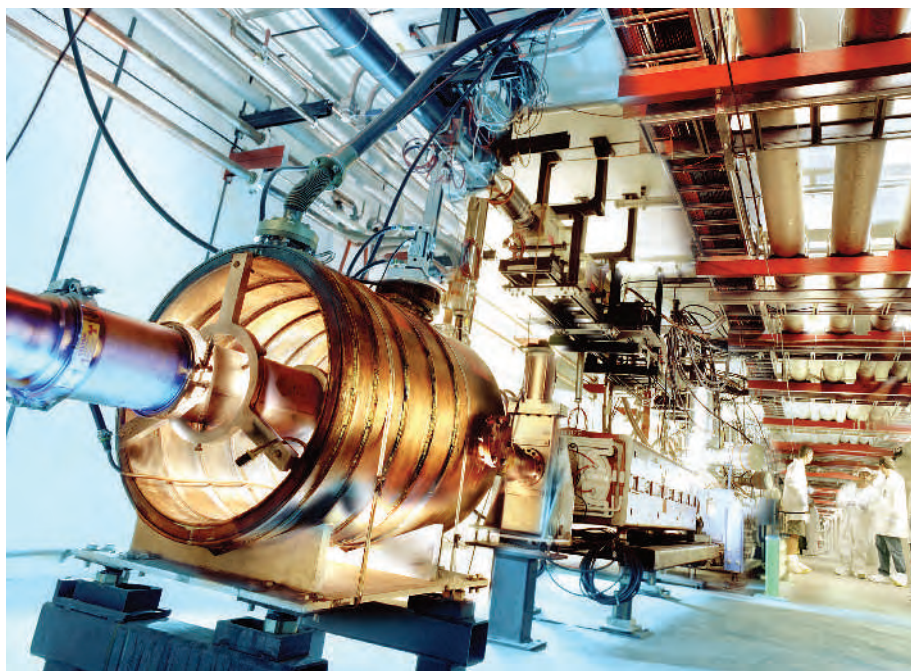
Keeping the faith

One solution may now be to diversify as SLAC has done. Fermilab has begun expanding its research in neutrinos — uncharged particles that pervade the Universe. Earlier this year, Fermilab turned on a \$170-million neutrino beam, aimed at a detector more than 700 kilometres away in the Soudan Mine in Minnesota. And several facilities are now being planned that would extend Fermilab's neutrino science well into the next decade.

But neutrino research will not be enough to sustain the facility at its present level, and many physicists at Fermilab remain disdainful of distractions from the lab's original mission. "Diversifying is the easier solution, but the hard road is more meaningful," says Young-Kee Kim, a physicist and spokesperson for the CDF. "From my point of view, we have to have a high-energy centre in the United States, or we'll lose the energy frontier."

Oddone agrees. "The International Linear Collider is the greatest opportunity for this lab right now," he says. With the US Department of Energy, Fermilab is stepping up its accelerator and detector research to make a strong bid for the new machine. If all goes well, Oddone says, the lab will again be the centre of world physics by early next decade as the linear collider probes exotic particles, such as the Higgs, in greater detail.

But that's a big if. Collaborations on the



FERMILAB PHOTO



Burton Richter (left) is critical of the lack of long-term planning that will put Fermilab's future in doubt after current machines (above) shut down.

don't want to work on something that might not start until 2016."

As a trip to the lab's cafeteria reveals, that problem is likely to get worse, not better. Sitting at one table is a group of four young Tevatron researchers from all over the world. They all say they support Fermilab's decision to build the linear collider, but only one — a Fermilab employee — says he will stay at the lab after CERN's machine is turned on.

The lab is doing its best to stem the tide. One stopgap is a computing centre that will allow US researchers to monitor their experiments at CERN from a distance. "We're going to see the same data as they see in CERN," says Avi Yagil, who is heading the centre's development. Fermilab is a critical link on the Grid, a distributed computing network that will transmit the Large Hadron Collider's data to the international physics community.

Yagil says he hopes that these efforts can prevent the lab from becoming, as he puts it, "scorched earth", but will it be enough to keep the lab going? Richter doesn't think so, and he is highly critical of the lab's leadership for not doing more to diversify their programme. "There has never been a coherent long-term view at Fermilab of what you're going to do in the next act, that is in 10 or 15 years," he says. "And that's what got them where they are."

Oddone begs to differ. Fermilab has the staff, knowledge and space to build the next big accelerator, he says. But anything less than a full commitment will not make the dream come true. "Yes, there are tremendous risks," he says. "But the questions we are trying to answer are also tremendous."

Geoff Brumfiel is *Nature's* Washington physical sciences correspondent.

"The International Linear Collider is the greatest opportunity for this lab right now." — Pier Oddone

scale needed to build the linear collider are notoriously difficult — fraught with cost overruns, infighting and international politics. And even if the collider does move forward, it may not be built in the United States. Japan is energetically pursuing its own plans to host the machine at its KEK facility, according to Yoji Totsuka, KEK's director-general.

The uncertainty is making it difficult for Fermilab to retain the thousands of visiting scientists it needs to run its existing experiments. Already, the Tevatron's two main detectors, CDF and DZero, are suffering from staff shortages. "A lot of people are in the process of moving or have already moved," says Jerry Blazey, spokesperson for DZero. "Right now we're just doing our best to hang on," adds Luciano Ristori, a CDF scientist.

This situation also makes it harder to attract the expertise needed to develop linear collider prototypes. "It's very difficult," says Hendrick Weerts, who has taken a sabbatical from Michigan State University in Lansing to work on prototype detectors. "Users know there's a really unique physics opportunity coming at CERN in the next two or three years, and they



Lie back and think of science

Twelve women have taken to bed for two months in the name of space research. **Nicola Jones** joins them for a few days, and asks what physiologists hope to learn from this marathon lie-in.

I'm staring at the ceiling and my feet are strapped to an exercise machine that acts like a giant yo-yo. My head rests a few centimetres lower than my feet, my face is bright red and my calves are screaming in pain as Björn Alkner, trained in physiology and physical education, yells at me to push the last set of repetitions to the maximum.

The exercise session ends and, unstrapped, I give in to the overwhelming desire to stand up and stretch. Instantly, I feel dizzy and my head pounds. Perhaps I'm not cut out for this. I've been here for just half an hour, but in the other rooms of this squat building at Toulouse's Rangueil Hospital lie women who have signed up to submit themselves to such procedures for two months — during which time they won't be able to sit or stand at all.

These volunteers will be pumped full of glucose, fed protein drinks, jabbed for blood samples and muscle biopsies, given electrical stimulation to make their muscles twitch, asked to run on a vertical treadmill, and much more. With no apparent irony, the researchers running this international project describe the experience as 'bedrest'.

They may seem like sadists, but the leaders of the project — run by MEDES, the French Institute for Space Medicine and Physiology in Toulouse — have a serious purpose. If astronauts are to spend longer stints in orbit, or

make the 30-month return trip to Mars, we need more data on how the body responds to microgravity — and how to mitigate its damaging effects on our physiology. Bedrest studies have for decades been used to mimic the weightlessness of space, but there is still much to learn; particularly about women, who have been under-represented in these experiments.

Difficult position

So the dozen volunteers in Toulouse are pioneers. "They are doing it for the sense of adventure," says MEDES project scientist Marie-Pierre Bareille. They certainly aren't in it for the money. Each volunteer will get €15,000 (US\$18,000) in instalments over four years. But it's a strange kind of adventure: the closest these women will get to the International Space Station (ISS) this summer is looking at its photo on the wall of the exercise room. For 60 days, they will lie with their heads at a 6° downwards slant, while they eat, shower...do everything, in fact. There are even discreet cameras in place to ensure that they don't cheat.

Despite this, the volunteers are enthusiastic. "The last time I thought about being an astronaut was when I was five years old," says the recumbent Dorota Tataruch, a 25-year-old student from Poland. "But all this is fascinating — talking to the scientists; seeing how we respond psychologically. I'm glad I'm here."

Lying down with your head lower than your feet may seem a poor imitation of the conditions in space, but it has many of the same physiological effects. After a few hours, blood and other fluids shift towards the head, making for a puffy face and headaches. After a day or so, the body responds to the increased volume of fluid in the upper torso: astronauts and bedrest volunteers quickly lose up to a litre of urine. Meanwhile, the lack of stimulation to balance sensors in the ears and nerves in the soles of the feet can cause dizziness and nausea. Astronauts can experience debilitating space sickness, and bedrest volunteers sometimes get a milder vertigo — usually relieved, for unknown reasons, by gently shaking the sufferer's head.

Leg and back muscles begin to atrophy — calf muscles can shrink by up to 30% within a few months, with a 50% reduction in some measures of leg strength. Every month, up to 2% of bone in the same areas is lost. As calcium leaches out, there's an increased risk of kidney stones. Without exercise, the body's control of blood glucose goes awry, potentially leading to insulin resistance. And without the effort of having to pump blood against gravity, the heart muscles shrink and 'forget' this crucial aspect of their function.

Even with two hours of exercise a day and a good diet, astronauts return to Earth severely weakened after a few months in space. "We

NASA

have some crew members who can't get out of their seats or walk," says Jeffrey Jones, head of NASA's medical operations for exploration missions at the Johnson Space Center in Houston, Texas.

Most bedrest studies to date have focused on changes in blood flow, bone and muscle mass. This study will also look at these effects, but with a few novel twists — including the gender make-up of the study group. Although there is no reason to think that female physiological responses to rest are radically different from men's, there are bound to be hormonal differences. And there are hints that women might faint more easily after a period in low gravity than men.

Undercover agents

The 12 women in the study have been split into three groups — one to explore the effects of exercise, one to test a high-protein diet, and the other serving as a control. But all face mental challenges. "Boredom and privacy are the first issues. Then you expect some to become short-tempered," says François-Régis Lenoir, a psychological counsellor on the trial. After 1,600 enquiries and 240 serious applications, only the most mentally and physically fit were selected. Still, there are bound to be difficulties. Tataruch's roommate had her 30th birthday while lying down, and all the team could do to celebrate was to wheel her outside for a bit of sunshine. At times, Tataruch and her nutrition-testing colleagues have been desperate to go for a run.

At least the exercise group has some sort of release. These women are testing two devices: the first is the yo-yo contraption I struggled with. This works like the squat-machines found in most gyms, although instead of lifting weights, you pull against the resistance of a nylon strap wrapped around a spindle. This exercises the thighs and calves, especially prone to muscle loss in space. The device performed well in a bedrest study with men, reportedly being both more effective and comfortable than the machines currently on the ISS — which include a stationary bike and a



Training station: astronauts must exercise in space to stop their leg muscles from wasting away.

treadmill that astronauts attach themselves to with bungee cords. "They're designing a model to test on the ISS now," says Alkner.

The second device is a vertical treadmill encased in a plastic box. The women are inserted into it sideways and air is extracted, which sucks the feet onto the jogging surface. This also pulls the blood to the feet, forcing the heart to work harder. It's an odd sensation being trussed up and swung into the device. Then the pump comes on and my feet hit the treadmill with a thud. It feels strange, but it works: now I'm running up the wall. Through about an hour of such exercises each day, researchers hope that the active group will retain 100% of their leg muscles — a feat never yet achieved.

The second experimental group is testing the idea that muscle loss can be minimized simply by eating a high-protein diet laced with extra leucine — an amino acid thought to be crucial to muscle building. "This should halve the amount of muscle loss," says Gianni Biolo from the University of Trieste in Italy, who is principal investigator on the project. The leucine comes in powdered

form. Added to water, it tastes foul.

Maintaining muscle mass and tone is one thing, but the effects of microgravity or bedrest on our nervous system and balance are harder to tackle. "That's our greatest weakness. It would be great if people focused more on that," says Jones. The Toulouse researchers will be looking at how some of our nerves and leg muscles work together through reflexes to help us walk without stumbling or falling over — using electrical stimulation, they will monitor how these reactions change with rest.

Bedside manner

Other project scientists are examining the effects of simulated microgravity on everything from the immune system to mental acuity. But the volunteers seem strangely subdued in the face of all this poking and prodding. They read, write e-mails or chat to each other. Counsellors and massage therapists are on hand to attend to any psychological and physical niggles.

When I call Tataruch from London to find out how her two months of 'rest' went, she is in good humour. "It passed much quicker than I expected," she says. Full results from the experiments won't be available for a while, but the diet and exercise regimes seem, from Tataruch's account, to have had the desired effect. Members of the control group managed to stand, albeit with a few wobbles, and painful backs and feet. Tataruch's leg muscles are visibly larger than theirs, she says, although smaller than those in the exercise group.

Now Tataruch faces a month of rehabilitation, to get back into sports and — she can't wait for this — real food. She says that the experience has prompted her to think about applying for astronaut training when she becomes eligible in a few years' time.

In the meantime, the same scientists are looking for more participants for bedrest. Any volunteers?

Nicola Jones is editor of news@nature.com.



Working flat out: your intrepid reporter tackles a prototype exercise machine for astronauts.

That sinking feeling

Siberia's vast forests absorb huge quantities of carbon from the atmosphere. But how much, and will they continue to do so in a warming world?

Quirin Schiermeier speaks to the carbon accountants.

It took just one Siberian heatwave to temporarily wipe out most of the gains made by the Kyoto Protocol on climate change. In the summer of 2003, wildfires raced across the region, incinerating an area of some 22 million hectares: slightly smaller than the state of Oregon. As the trees went up in smoke, they released about 250 million tonnes of carbon into the atmosphere — roughly the same amount as industrialized countries have pledged to cut from their emissions by 2012 under the Kyoto agreement.

Natural releases of carbon — from wild fires, thawing permafrost or drained peatlands — aren't covered by the Kyoto Protocol, which aims to cut emissions from the burning of fossil fuels. But if future political agreements are to save us from a climate catastrophe, experts agree that we need a much better understanding of natural carbon sources and sinks, and whether they can be manipulated to help put a brake on global warming.

Siberia looms large in this debate, its vast forests — or taiga — are a potentially huge sink for greenhouse gases such as carbon dioxide. Yet so little is known about Siberia's role in the global carbon cycle that researchers have been scrambling to gather basic data. For the past few years, scientists have fanned out across the taiga, working in two projects aimed at quantifying Siberia's carbon budget.

"We need to understand what will happen when climate change really kicks off," says Wolfgang Lucht, a biosphere researcher at the Potsdam Institute for Climate Impact Research in Germany. "When, in 30 years, governments ask us how the world has changed, we must be able to provide answers."

To that end, 14 institutes have taken part in the Siberia-II, a European Union-funded

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS



Taiga, taiga, burning bright: following huge forest fires in Siberia (above), researchers have been monitoring the amount of greenhouse gases released by the trees (left).

M. HEIMANN

initiative to quantify the greenhouse-gas budget of a 2-million-square-kilometre area in central Siberia. A complementary project, called the Terrestrial Carbon Observing System, or TCOS, is providing more detailed, real-time measurements of carbon fluxes between Siberian forests and the lower levels of the atmosphere. Both projects, now approaching completion, have yielded a wealth of data. And one fact stands out: Siberia's taiga is a more modest carbon sink than previously thought.

Lie of the land

These vast forests are by far the largest terrestrial carbon reservoir in the Northern Hemisphere. Unlike tropical rain forests, where logging is the main disturbance to the carbon cycle, changes in the carbon balance in the taiga are being driven mainly by climate change. This makes these forests an ideal place to study the interplay between climate and landscape.

Siberia is also a climatic hot spot, with short warm summers and extremely long, cold winters. Temperatures vary greatly from year to year, but on average, surface temperatures in central Siberia have increased by up to 3 °C since 1960 — three to four times more than the global average.

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

M. O. THIESSEN/NATL GEOGRAPHIC/GETTY IMAGES

In a warmed world of the future, Siberia might see some short-term benefits. The warmer temperatures have already sparked a greening trend: spring arrives sooner and buds and leaves appear earlier in the growing season. To some extent, this counteracts the greenhouse effect, as growing plants remove carbon dioxide from the atmosphere through photosynthesis. Some Russian scientists and policy-makers critical of the Kyoto Protocol claim that a warmer climate will have positive effects on plant productivity and agriculture in Russia¹.

Tinderbox

But the growth spurt is unlikely to continue for ever, says Lucht. "At some point, even though we don't know when, Siberian forests might begin to suffer from heat stress and die," he explains. And the long-term picture isn't much better. Although the natural fluctuations are large, and model predictions uncertain, many scientists believe that summers will become warmer and drier in Siberia. This would favour the outbreak of fires, releasing carbon at an even greater rate, says Heiko Baltzer, head of Earth observation at the Centre for Ecology and Hydrology in Monks Wood, UK.

Recent research shows that a warmer climate could also enhance soil respiration, permafrost thawing and the decomposition of organic matter in peat, moors and bogs — all of which result in additional release of carbon^{2,3}. By the middle of this century, the

planet's terrestrial biosphere could turn from a carbon sink into a carbon source⁴ — and global warming would step up a gear.

Siberia-II has tried to document the current state of carbon exchange in the region's forests so that future changes can be carefully measured. But it hasn't been an easy task. Project scientists have learned that at least one crucial bit of information is missing: the total biomass contained in Siberia's forests and soils. Satellites can monitor changes in land cover, but they provide little or no information about the actual volume of stems, roots and leaves present.

Lost in a forest

So forest inventories are crucial. Unfortunately, they are scarce for Siberia, and those that do exist are full of gaps. "Most data from Russian inventories are obsolete, and some are plain wrong," says Anatoly Shvidenko, a forest ecologist with the International Institute for Applied Systems Analysis (IIASA) in Laxenburg, Austria. "For example, official Russian statistics either don't account for the large forest fires in 1998 and 2003, or substantially underestimate them," he says.

But Shvidenko, former head of the All-Russian Scientific Research Information Center for Forest Resources in Moscow, is an old hand at dealing with Russian sources. With the help of two partner institutes in Irkutsk and Krasnojarsk, he has touted round and collected together what he believes are the most accurate data on forest inventories, discarding information he knew to be outdated or misleading. This newly compiled inventory will be indispensable for validating remote-sensing data and the results of computer models.

The Siberia-II researchers have also developed mathematical models to transform raw satellite data into real-world information, such as how much of the taiga is covered by plants or snow. This information has been checked by scientists heading out into the field to confirm that the satellites were correctly interpreting what they saw on the ground. In addition, team members have produced new maps, which can be used to improve the accuracy of computer models about vegetation in Siberia.

The TCOS project, meanwhile, focused on the flux of carbon dioxide between the forest and the atmosphere. Six observational towers, each placed in a separate ecosystem, measured gases in the air up to a height of 30 metres. Once or twice a month, scientists flew aircraft through the lower 2 kilometres of the atmosphere, to measure carbon dioxide levels.

Early results have shown not only that the taiga is a relatively modest sink for carbon, but that it may also give off more methane — a strong greenhouse gas — than previously thought. Both effects have yet to be fully quan-

tified, but the two projects have made scientists more confident that they can account for the carbon budget for an area this large.

The preliminary findings have whetted scientists' appetite for more. In particular, Christiane Schmullius, a remote-sensing expert at Jena University in Germany, who coordinated Siberia-II, wants to use radar to monitor biomass from space. The European Space Agency's Envisat satellite carries a radar sensor, but its ability to determine the height and density of vegetation is limited.

"We have seen in Siberia that remote sensing has the potential to provide all the information needed for full greenhouse-gas accounting," says Schmullius. NASA plans to take advantage of that fact for its Orbiting Carbon Observatory mission, set to launch in 2008. This satellite will make global carbon dioxide maps twice a month, at a level of detail that would allow it to monitor changes such as those triggered by Siberian wildfires.

Branching out

Siberia-II has substantially helped to reduce the uncertainties about carbon flux, says Sten Nilsson, a forest scientist at the IIASA. "Uncertainties used to be in the range of plus or minus 100%," he says. "Now we may have halved the margin of error."

Buoyed by the early results, Germany's Max Planck Society has funded an ambitious follow-up project to establish in greater detail how greenhouse gases flow between Siberian forests and the atmosphere. Together with Russian partners, the society is building a 300-metre-high measurement tower near the Yenisey River in central Siberia. The Zotino Tall Tower Observatory, or Zotto, will become operational next year, fully equipped with high-precision tools to measure a variety of greenhouse

gases. Zotto will also have an underground laboratory, buried to protect it from climate extremes, for *in situ* analysis.

Zotto should help constrain assumptions on how much more greenhouse gas the taiga is capable of sucking up, says Martin Heimann, an atmospheric physicist at the Max Planck Institute for Biogeochemistry in Jena, who coordinates TCOS. Such information, accumulated piece by piece, helps to flesh out the full picture of carbon accounting and so could shape the direction of future climate policies. It may prove to be possible to sort the wood from the trees after all. ■

Quirin Schiermeier is *Nature's* German correspondent.

"Uncertainties about the carbon flux used to be in the range of plus or minus 100% — now we may have halved the margin of error."
— Sten Nilsson

1. Schiermeier, Q. & MacWilliams, B. *Nature* **431**, 12–13 (2004).
2. Frey, K. E. & Smith, L. C. *Geophys. Res. Lett.* **32**, L09401 (2005).
3. Knorr, W., Prentice, I. C., House, J. I. & Holland, E. A. *Nature* **433**, 298–301 (2005).
4. Cox, P. M., Betts, R. A., Jones, C. D., Spall, S. A. & Totterdell, I. J. *Nature* **408**, 184–187 (2000).

BUSINESS

Touched by an angel

Looking for a fairy godmother to back a brilliant business idea? As **Tony Reichhardt** reports, a group of them probably meets for breakfast once a month at a hotel near you.

Kevin Short could have stayed a mathematician. Studying nonlinear dynamics at the University of New Hampshire in Durham had its pleasures, but also its limitations. As a young man working at IBM, he'd observed how the older managers came alive when they talked about building the first satellites in the 1960s. "They just glowed," he recalls. Short wanted to see his own ideas for applying chaos mathematics to telecommunications become reality. But he knew that a university mathematics department was not the place to do it.

So in 2000 Short did what many a scientist with a 'big idea' has done. He started a company. And like many new company owners, his escort into the world of finance was an 'angel', a small investor who steps in after you've cadged your last penny from the 'three Fs' — friends, family and fools — but before you need the larger sums that venture-capital firms supply.

Angels don't put up as much money as the venture capitalists, but they are involved in far more deals — some 48,000 last year in the United States, with a total value of \$22.5 billion, according to the University of New Hampshire's Center for Venture Research. By one estimate, angels fund 60% of all high-tech firms in the United States.

The average deal is in the hundreds of thousands of dollars. Short's Massachusetts-based company Groove Mobile (formerly Chaoti-

com) is competing in the hot market to download music to mobile phones. He initially got \$500,000 from a group of five or six angels. Computer-industry executive Dave Croston needed about the same amount to get his online-data security company IAM Registry of North Hampton, New Hampshire, off the ground last year. Co-founding the company with two computer scientists from Brown University and the University of California, Irvine, Croston says that "a little money is far more difficult to raise than a lot of money". So "you deal with angels completely differently than you deal with venture capitalists".

Emotional ties

Whereas venture-capital firms tend to be somewhat formal, Croston says, "the angel is a very finicky, emotional character. And that's not a bad thing." Angels tend to be hands on and they tend to buy into businesses whose product they understand well. Jeffrey Sohl of the Center for Venture Research says that 90% live within a day's drive of their investments. They favour high-tech fields such as software, medical equipment and biotechnology.

And the entrepreneur often gets a lot more from the deal than just money. Kevin Short's chief angel was 62-year-old George McQuilken, who followed a long career at IBM by starting several computer-security businesses and then co-founded New Hampshire's eCoast Angel Network in 2000. As well as investing his own money, McQuilken had useful contacts and could advise on business plans. With his background in information theory, he could even show sceptics the door. When one challenged Short's technology on the grounds that it violated Shannon's law, which governs how far data can be compressed, McQuilken could argue from experience. "I knew Claude Shannon," he says.

Croston advises would-be entrepreneurs that "just finding a rich guy isn't enough. Find somebody in your field". An angel who understands the technology and also understands business can help convert your original idea into something marketable. "Very seldom

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Angels provide cash when start-ups need it.

do you turn out to be who you think you're going to be," says Croston. "You start off as a square box, and end up as a round hatbox." It was McQuilken's daughter Lucy, herself a veteran of the semiconductor industry, who helped steer Short's Chaoticom into the downloadable-music market.

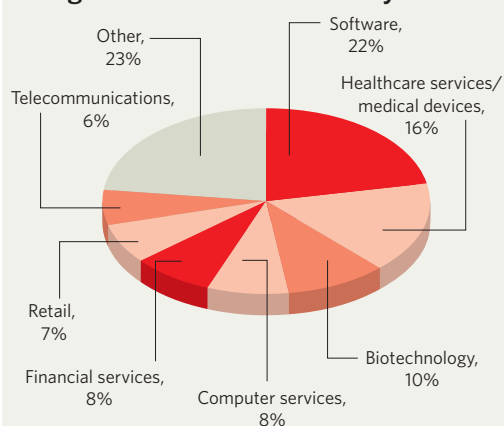
Scientists dealing with angels need to learn humility and to leave business to the experts, says Sohl. Investors "like the scientist who doesn't want to be the chief executive," he says. Short soon learned that business and research are "different skill sets". Just because an idea is intellectually elegant doesn't mean it will pay off.

Money pooled

Angel investors are now starting to band together, forming small networks such as Silicon Valley's Band of Angels. McQuilken's eCoast Angels started with 16 people about five years ago. Back then, it was all individual deals — "parallel play," he says. Now the members often pool their investments, relying on each other for advice and expertise in technical fields outside their own. New members frequently sit back and watch at first. "It might take them a year to write their first cheque." And there's no particular pressure to go faster. The eCoast Angels only buy into two or three new deals a year.

The latest trend is for angel networks to form even larger, but still loosely organized, groups such as the Angel Capital Association based in Kansas City, Missouri, the European Business Angel Network (EBAN) and the Business Angels Network Southeast Asia.

Angel investments in 2004 by sector



SOURCE: CENTER FOR VENTURE RESEARCH

J. LEVINE/CORBIS

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

According to EBAN there were approximately 280 business-angel networks in Europe last year, with a total investment estimated at one-tenth that of US angels. The ultimate aim of all these associations and *uber*-associations is to act as dating services to match entrepreneurs with the perfect angel. They also try to raise the visibility of the field, and to lobby governments for tax breaks for small investors.

Sohl estimates that there are currently about 225,000 active angels in the United States. Like the larger venture-capital firms, angels got slammed by market conditions in the early part of this decade. But like the big boys, they're starting to come back. The number of deals in 2004 was up 24% from the previous year, Sohl estimates.

It is, of course, ultimately about the money. Everyone dreams of a golden investment like the Body Shop, the natural beauty products business that reportedly brought some lucky angel 10,500 times their initial stake. On the basis of his own informal accounting, however, McQuilken thinks a 30% return is more typical.

But he and others say that many angels see a reward in building up local businesses along with the financial payoff. The money they put into local ventures — which sometimes benefit university labs as well — has a significant impact on regional economies.

And a local view works well for the angels too. "There's probably no major venture capitalist in Boston who has the faintest thought of investing in Keene, New Hampshire. They're investing in India," McQuilken says. As a result, "We look at a lot of deals that aren't shopworn."

Tony Reichhardt

IN BRIEF

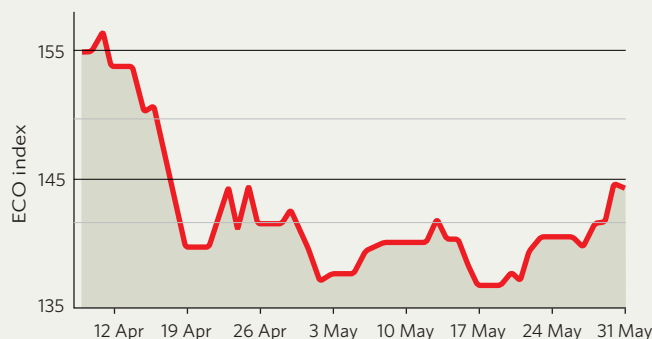
GOING FOR A TWIRL Helicopter manufacturer Sikorsky says it is planning to build a prototype helicopter that will make use of new materials, electronics and other systems to fly one-and-a-half times as fast as any existing model. The Connecticut-based company says that the craft will use advanced composite materials, an active vibration control system, and two rotor blades turning in opposite directions on the same axis, to propel itself at almost 290 kilometres per hour. It could fly by the end of next year.

DOUBLE CHIPS Advanced Micro Devices (AMD), the firm in Sunnyvale, California, best known for selling cheap, Intel look-alike microchips, has stolen a march on its more famous rival by launching the first, genuine 'dual-core' microprocessor for use in personal computers. The AMD Athlon 64 X2 chip, which starts at \$500, raises performance by integrating two processors on the same chip. Analysts say that dual-core chips will enable manufacturers to build ever-faster chips that don't demand outrageous amounts of power and cooling.

VIAGRA WARNING The US Food and Drug Administration has asked Pfizer to modify the warning label on the firm's anti-impotence drug Viagra while the agency investigates whether 43 reported cases of sudden blindness in men taking impotence drugs are coincidental or linked to the medications. Thirty-eight of the cases occurred in men taking Viagra. The company said, however, that "there is no evidence showing that [blindness] occurred more frequently in men taking Viagra than men of similar age and health who did not take Viagra."

MARKET WATCH

Clean energy stocks



Companies in the renewable-energy sector saw their market valuations slip back in April, and then regain some ground last month.

The WilderHill Clean Energy Index (ECO on the American Stock Exchange; see graph) reflects the subsectors that make up the renewable-energy industry. It was set up six years ago by Robert Wilder, at the time a political scientist at the University of California, Santa Barbara. He left to form WilderShares of Encinitas, California, which runs the index and sells a fund made up of the quoted stocks.

"We've focused not on the largest companies, but on those with the most significant technologies," Wilder says. Suppliers of storage devices and superconducting transmission cables, as well as solar panels and windmills, are included, as are utilities — such as Scottish Power — with strong

renewable-energy interests. Nuclear power is shut out.

Typical companies are Emcore of New Jersey, a semiconductor maker that produces unusually efficient photovoltaic cells, and Evergreen Solar, a Massachusetts firm whose 'string ribbon' technology reduces the amount of crystalline silicon needed to make such cells. The companies are based in several countries but all are quoted on the American Stock Exchange or on NASDAQ.

Wilder says that the index's decline earlier this year was accentuated by profit-taking in solar- and wind-power stocks, whose prices had been inflated earlier by strong product demand, especially in Germany and Japan.

He is bullish about its long-term prospects. Renewable energy "is becoming acceptable, even to conservative companies", he says.

SOURCE: WILDERSHARES

Controls on exports will defend security, without harming labs

SIR — Your News story “Academics stress licence threat to US science” (*Nature* 435, 4; 2005) usefully focuses attention on US export-control regulations and their application to scientific research. But it repeats a common misperception regarding the scope of the export-control regulations in the context of domestic academic laboratories.

The News story states that a licence is required for certain foreign nationals to operate controlled equipment. This is not the case. A foreign national may work with controlled equipment, so long as no controlled technology is released to the foreign national in the process. The controls apply to the release of technology, not to the use of equipment.

The News story also states that the controls apply to technologies relating to a wide range of equipment. But licences are only required for release of a relatively small subset of technologies to foreign nationals from a limited number of countries. For instance, the article cites lasers as an example, yet no licences would be required to teach a foreign national how to use most lasers.

Finally, US companies and many universities are already complying with these export controls. The president of the University of Maryland, Dan Mote, is quoted as considering that obtaining the required licences could “bring work to a halt” in our nation’s laboratories. On the contrary, proper compliance with these rules will strengthen our nation’s security while ensuring that American science maintains its pre-eminent research capabilities.

Peter Lichtenbaum

Acting undersecretary for industry and security,
US Department of Commerce,
14th Street and Pennsylvania Avenue,
Washington DC 20230, USA

Veracity of raw images can also come into question

SIR — Your News Feature “CSI: cell biology” (*Nature* 434, 952–953; 2005) raises questions about the legitimacy of editing images. In this context, I am regularly surprised at the essentially binary appearance of many published fluorescence microscopy images, with the apparent absence of both intermediate intensities and Poisson noise.

This may predate Photoshop and arise at the time of acquisition, when alterations to the offset of photodetectors can be used to suppress low or intermediate signals and

the gain setting used to produce detector saturation by the remaining portion of the original signal.

Although attempting to curtail illegitimate manipulation of digital images is very important, it is also relevant to consider the veracity of raw images. A solution would be to require a statement about detector settings in the methods section of publications and, further, to require that original unenhanced digital images, in the proprietary format, which will usually contain details of the microscope settings, are included in online supporting material.

Jeremy Adler

Department of Developmental Biology,
The Wenner-Gren Institute,
Stockholm University, S106 91,
Stockholm, Sweden

Academic boycott would damage chances for peace

SIR — As your News story, “Palestinian unease sparks fresh calls for Israeli boycott” (*Nature* 434, 813; 2005) notes, intermittent attempts have been made to boycott Israeli academics, in order to protest at the lack of resolution of the Israeli–Arab conflict.

These efforts — such as the recent attempt by some members of the UK Association of University Teachers to impose a boycott of two universities, which was overturned last month (*Nature* 435, 550; 2005) — are counterproductive in an environment where mistrust is a key obstacle to peace, particularly after four years of escalated violence. This has included closures of Palestinian universities by Israeli forces and direct attacks on Israeli institutions of higher learning, such as the terrorist bombing of the cafeteria at Hebrew University that killed and injured nearly 100 people, mostly students.

In addition to reducing the enmity that must be overcome before peace can be achieved, collaboration between Israeli and Palestinian academics has brought millions of dollars in funding to suburban Jerusalem’s Al-Quds University. Joint efforts have enhanced the training and skills of the Palestinian faculty members — skills they and their students will need in any future Palestinian state, if it is to be viable, productive and at peace with its neighbours.

Even at the height of the Cold War, with nuclear missiles aimed at each other’s heartlands, the United States and the Soviet Union considered academic exchanges a worthwhile way of enhancing the prospects for peace. Although there are, of course, major differences between the two conflicts, the same principle holds true for Israelis and Palestinians. Only those who are interested in perpetuating this conflict indefinitely

stand to gain anything from preventing such exchanges.

John R. Cohn

Thomas Jefferson University, 1015 Chestnut
Street, Philadelphia, Pennsylvania 19107, USA

A boycott could do good in Israel, as in South Africa

SIR — We find it curious that your News story “Palestinian unease sparks fresh calls for Israeli boycott” (*Nature* 434, 813; 2005) portrays scientific collaboration between Palestinian and Israeli academic institutions as a “beacon of hope” at a time when Israel’s oppression of the Palestinians is ongoing and, in fact, intensifying.

Plain logic would lead us to think that liberation is required first, and then perhaps collaboration, with justice having already been achieved. To suggest that collaboration despite occupation is a sign of hope defies any understanding of conflict, oppression and the struggle for freedom, justice and genuine peace.

In our view, boycott constitutes one of the very few possibilities for Palestinian non-violent resistance to occupation. Boycott as an instrument of civil resistance did not originate in Palestine. It has been effectively used elsewhere, notably in South Africa, and has earned much support from various groups worldwide.

In the case of Palestine–Israel, a moral, if not active, support of the boycott of Israeli academic and cultural institutions is the least that one expects from conscientious academics, scientists and intellectuals worldwide. Would *Nature* have described the collaboration of blacks and other liberationists in South Africa with the apartheid government and its institutions as a “beacon of hope”?

Rita Giacaman*, Jacqueline Sfeir†, Ismat al-Shakhshir‡

*Institute of Community and Public Health,
Birzeit University, Birzeit, West Bank, Palestine

†Department of Education, Bethlehem University,
Bethlehem, West Bank, Palestine

‡Department of Chemistry, Al-Najah University,
Nablus, West Bank, Palestine

The News story in question was describing events and attitudes, not presenting an editorial perspective. For the record, *Nature* has consistently opposed boycotts, both of South African academics in the time of apartheid and, more recently, of researchers in Israel — Editor, *Nature*.

Contributions to Correspondence may be submitted to corres@nature.com. They should be no longer than 500 words, and ideally shorter. Published contributions are edited.

COMMENTARY

Scientists behaving badly

To protect the integrity of science, we must look beyond falsification, fabrication and plagiarism, to a wider range of questionable research practices, argue **Brian C. Martinson, Melissa S. Anderson and Raymond de Vries.**

Serious misbehaviour in research is important for many reasons, not least because it damages the reputation of, and undermines public support for, science. Historically, professionals and the public have focused on headline-grabbing cases of scientific misconduct, but we believe that researchers can no longer afford to ignore a wider range of questionable behaviour that threatens the integrity of science.

We surveyed several thousand early- and mid-career scientists, who are based in the United States and funded by the National Institutes of Health (NIH), and asked them to report their own behaviours. Our findings reveal a range of questionable practices that are striking in their breadth and prevalence (Table 1). This is the first time such behaviours have been analysed quantitatively, so we cannot know whether the current situation has always been the case or whether the challenges of doing science today create new stresses. Nevertheless, our evidence suggests that mundane 'regular' misbehaviours present greater threats to the scientific enterprise than those caused by high-profile misconduct cases such as fraud.

As recently as December 2000, the US Office of Science and Technology Policy (OSTP) defined research misconduct as "fabrication, falsification, or plagiarism (FFP) in proposing, performing, or reviewing research, or in reporting research results". In 2002, the Federation of American Societies for Experimental Biology and the Association of American Medical Colleges objected to a proposal by the US Office of Research Integrity (ORI) to conduct a survey that would collect empirical evidence of behaviours that can undermine research integrity, but which fall outside the OSTP's narrow definition of misconduct^{2,3}. We believe that a valuable opportunity was wasted as a result.

A proper understanding of misbehaviour requires that attention be given to the negative aspects of the research environment. The modern scientist faces intense competition, and is further burdened by difficult, sometimes unreasonable, regulatory, social, and managerial demands⁴. This mix of pressures creates many possibilities for

"Our findings suggest that US scientists engage in a range of behaviours extending far beyond falsification, fabrication and plagiarism."

Table 1 | Percentage of scientists who say that they engaged in the behaviour listed within the previous three years ($n = 3,247$)

Top ten behaviours	All	Mid-career	Early-career
1. Falsifying or 'cooking' research data	0.3	0.2	0.5
2. Ignoring major aspects of human-subject requirements	0.3	0.3	0.4
3. Not properly disclosing involvement in firms whose products are based on one's own research	0.3	0.4	0.3
4. Relationships with students, research subjects or clients that may be interpreted as questionable	1.4	1.3	1.4
5. Using another's ideas without obtaining permission or giving due credit	1.4	1.7	1.0
6. Unauthorized use of confidential information in connection with one's own research	1.7	2.4	0.8 ***
7. Failing to present data that contradict one's own previous research	6.0	6.5	5.3
8. Circumventing certain minor aspects of human-subject requirements	7.6	9.0	6.0 **
9. Overlooking others' use of flawed data or questionable interpretation of data	12.5	12.2	12.8
10. Changing the design, methodology or results of a study in response to pressure from a funding source	15.5	20.6	9.5 ***
Other behaviours			
11. Publishing the same data or results in two or more publications	4.7	5.9	3.4 **
12. Inappropriately assigning authorship credit	10.0	12.3	7.4 ***
13. Withholding details of methodology or results in papers or proposals	10.8	12.4	8.9 **
14. Using inadequate or inappropriate research designs	13.5	14.6	12.2
15. Dropping observations or data points from analyses based on a gut feeling that they were inaccurate	15.3	14.3	16.5
16. Inadequate record keeping related to research projects	27.5	27.7	27.3

Note: significance of χ^2 tests of differences between mid- and early-career scientists are noted by ** ($P < 0.01$) and *** ($P < 0.001$).

the compromise of scientific integrity that extend well beyond FFP.

We are not the first to call attention to these issues — debates have been ongoing since questionable research practices and scientific integrity were linked in 1992 report by the National Academy of Sciences⁵. But we are the first to provide empirical evidence based on self reports from large and representative samples of US scientists that document the occurrence of a broad range of misbehaviours.

The few empirical studies that have explored misbehaviour among scientists rely on confirmed cases of misconduct⁶ or on scientists' perceptions of colleagues' behaviour⁷⁻⁹, or have used small, non-representative samples of respondents^{8,9}. Although inconclusive, previous estimates of the prevalence of FFP range from 1% to 2%. Our 2002 survey was based on large, random samples of scientists drawn from two databases that are maintained by the NIH Office of

Extramural Research. The mid-career sample of 3,600 scientists received their first research-project (R01) grant between 1999 and 2001. The early-career sample of 4,160 NIH-supported postdoctoral trainees received either individual (F32) or institutional (T32) postdoctoral training during 2000 or 2001.

Getting data

To assure anonymity, the survey responses were never linked to respondents' identities. Of the 3,600 surveys mailed to mid-career scientists, 3,409 were deliverable and 1,768 yielded usable data, giving a 52% response rate. Of the 4,160 surveys sent to early-career scientists, 3,475 were deliverable, yielding 1,479 usable responses, a response rate of 43%.

Our response rates are comparable to those of other mail-based surveys of professional populations (such as a 54% mean response rate from physicians¹⁰). But our approach certainly leaves room for potential non-response bias; misbehaving scientists may have been less likely than others to respond to our survey, perhaps for fear of discovery and potential sanction. This, combined with the fact that there is

probably some under-reporting of misbehaviours among respondents, would suggest that our estimates of misbehaviour are conservative.

Our survey was carried out independently of, but at around the same time as, the ORI proposal. The specific behaviours we chose to examine arose from six focus-group discussions held with 51 scientists from several top-tier research universities, who told us which misbehaviours were of greatest concern to them. The scientists expressed concern about a broad range of specific, sanctionable conducts that may affect the integrity of research.

To affirm the serious nature of the behaviours included in the survey, and to separate potentially sanctionable offences from less serious behaviours, we consulted six compliance officers at five major research universities and one independent research organization in the United States. We asked these compliance officers to assess the likelihood that each behaviour, if discovered, would get a scientist into trouble at the institutional or federal level. The first ten behaviours listed in Table 1 were seen as the most serious: all the officers judged them as likely to be sanctionable, and at least four of the six officers judged them as very likely to be sanctionable. Among the other behaviours are several that may best be classified as carelessness (behaviours 14 to 16).

Admitting to misconduct

Survey respondents were asked to report in each case whether or not ('yes' or 'no') they themselves had engaged in the specified behaviour during the past three years. Table 1 reports the percentages of respondents who said they had engaged in each behaviour. For six of the behaviours, reported frequencies are under 2%, including falsification (behaviour 1) and plagiarism (behaviour 5). This finding is consistent with previous estimates derived from less robust evidence about misconduct. However, the frequencies for the remaining behaviours are 5% or above; most exceed 10%. Overall, 33% of the respondents said they had engaged in at least one of the top ten behaviours during the previous three years.

Among mid-career respondents, this proportion was 38%; in the early-career group, it was 28%. This is a significant difference ($\chi^2 = 36.34$, d.f. = 1, $P < 0.001$). For each behaviour where mid- and early-career scientists' percentages differ significantly, the former are higher than the latter.

Although we can only speculate about the observed sub-group differences, several explanations are plausible. For example, opportunities to misbehave, and perceptions of the likelihood or consequences of being caught, may change during a scientist's career. Or it may be that these groups

received their education, training, and work experience in eras that had different behavioural standards. The mid-career respondents are, on average, nine years older than their early-career counterparts (44 compared with 35 years) and have held doctoral degrees for nine years longer.

Another possible explanation for sub-group differences is the under-reporting of misbehaviours by those in relatively tenuous, early-career positions. Over half (51%) of the mid-career respondents have positions at the associate-professor level or above, whereas 58% of our early-career sample are post-doctoral fellows.

Addressing the problem

Our findings suggest that US scientists engage in a range of behaviours extending far beyond FFP that can damage the integrity of science. Attempts to foster integrity that focus only on FFP therefore miss a great deal. We assume that our reliance on self-reports of behaviour is likely to lead to under-reporting and therefore to conservative estimates, despite assurances of anonymity. With as many as 33% of our survey respondents admitting to one or more of the top-ten behaviours, the scientific community can no longer remain complacent about such misbehaviour.

Early approaches to scientific misconduct focused on 'bad apples'. Consequently, analyses of misbehaviour were limited to discussions of individual traits and local (laboratory and departmental) contexts as the most likely determinants. The 1992 academy report⁵ helped shift attention from individuals with 'bad traits' towards general scientific integrity and the 'responsible conduct of research.'

Over the past decade, government agencies and professional associations interested in promoting integrity have focused on responsible conduct in research^{5,11,12}. However, these efforts still prioritize the immediate laboratory and departmental contexts of scientists' work, and are typically confined to 'fixing' the behaviour of individuals.

Missing from current analyses of scientific integrity is a consideration of the wider research environment,

including institutional and systemic structures. A 2002 report from the Institute of Medicine directed attention to the environments in which scientists work, and recommended an institutional (primarily university-level) approach to promoting responsible research¹³. The institute's report also noted the potential importance of the broader scientific environment, including regulatory and funding agencies, and the peer-review system, in fostering or hindering integrity, but remained mostly silent on this issue owing to a dearth of evidence.

In our view, certain features of the research working environment may have unexpected

and potentially detrimental effects on the ethical dimensions of scientists' work. In particular, we are concerned about scientists' perceptions of the functioning of resource distribution processes. These processes are embodied in professional societies, through peer-review systems and other features of the funding and publishing environment, and through markets for research positions, graduate students, journal pages and grants. In ongoing analyses, not yet published, we find significant associations between scientific misbehaviour and perceptions of inequities in the resource distribution processes in science. We believe that acknowledging the existence of such perceptions and recognizing that they may negatively affect scientists' behaviours will help in the search for new ways to promote integrity in science.

Little attention has so far been paid to the role of the broader research environment in compromising scientific integrity. It is now time for the scientific community to consider what aspects of this environment are most salient to research integrity, which aspects are most amenable to change, and what changes are likely to be the most fruitful in ensuring integrity in science. ■

Brian C. Martinson is at the HealthPartners Research Foundation, 8100 34th Avenue South, PO Box 1524, Mailstop 2111R, Minneapolis, Minnesota 55440-1524, USA.

Melissa S. Anderson is at the University of Minnesota, Educational Policy and Administration, 330 Wulling Hall, Minneapolis, Minnesota 55455, USA.

Raymond de Vries is at the University of Minnesota, Center for Bioethics, N504 Boynton, Minneapolis, Minnesota 55455, USA.

1. OSTP *Federal Policy on Research Misconduct* http://www.ostp.gov/html/001207_3.html (2005).
2. Teitelbaum, S. L. *Nature* **420**, 739–740 (2002).
3. Korn, D. *Nature* **420**, 739 (2002).
4. Freeman, R., Weinstein, E., Marincola, E., Rosenbaum, J. & Solomon, F. *Science* **294**, 2293–2294 (2001).
5. *Panel on Scientific Responsibility and the Conduct of Research* (Natl Acad., Washington DC, 1992).
6. Steneck, N. H. *ORI Introduction to the Responsible Conduct of Research* (US Government Printing Office, Washington DC, 2004).
7. Swazey, J. M., Anderson, M. S. & Louis, K. S. *Am. Sci.* **81**, 542–553 (1993).
8. Ranstam, J. et al. *Control Clin. Trials* **21**, 415–427 (2000).
9. Geggie, D. *J. Med. Ethics* **27**, 344–346 (2001).
10. Asch, D. A., Jedrzejewski, M. K. & Christakis, N. A. *J. Clin. Epidemiol.* **50**, 1129–1136 (1997).
11. *Committee on Science Engineering and Public Policy On Being a Scientist: Responsible Conduct in Research* (Natl Acad., Washington DC, 1995).
12. *Panel on Scientific Responsibility and the Conduct of Research* (Natl Acad., Washington DC, 1993).
13. Institute of Medicine and National Research Council *Committee on Assessing Integrity in Research Environments: Integrity in Scientific Research: Creating an Environment that Promotes Responsible Conduct* (Natl Acad., Washington DC, 2002).

Acknowledgements: This research was supported by the Research on Research Integrity Program, an ORI/NIH collaboration, with financial support from the National Institute of Nursing Research and an NIH Mentored Research Scientist Award to R.d.V. We thank the three anonymous reviewers, Nick N. Steneck and M. Sheetz for their insightful input and responses to earlier drafts.

BOOKS & ARTS

Born to greatness?

Staking a claim in history for one of the driving forces behind quantum mechanics.

The End of the Certain World: The Life and Science of Max Born

by Nancy Thorndike Greenspan
Basic Books: 2005. 320 pp. \$26.95

Kurt Gottfried

Max Born was one of the founding fathers of quantum mechanics — indeed, he coined its name even before his assistant, Werner Heisenberg, gave birth to the theory with a breakthrough paper in the summer of 1925. But Born never received the recognition he so richly deserved, a gap that this book should help to fill.

Born was the professor of theoretical physics in Göttingen from 1921 until his dismissal by the Nazis in 1933. In those days that meant he could really shape the content and quality of the research programme. This he did superbly, assembling a group of brilliant youngsters and focusing them on atomic physics.

By 1925, Born had already realized that the puzzles posed by atomic spectra should be attacked by focusing on transition probabilities. He had also reached the seminal view that the theory should concentrate on what is in principle observable, and not on classical but unobservable constructs, such as electronic orbits, which figured prominently in the 'old quantum theory'. These insights were essential conceptual ingredients in Heisenberg's paper, which first set out, albeit in a skeletal and rather opaque form, the algebraic scheme of the nascent theory. It was Born who recognized that this scheme was matrix algebra. And it was Born who first wrote down the commutation rule, which specifies by how much xp differs from px , where x and p are the matrices corresponding to the coordinate and momentum of a particle. This equation is the key to the mathematical structure of quantum mechanics, and was engraved on Born's tombstone at his request.

Before the end of 1925, Born, Heisenberg and Pascual Jordan had developed these first steps at amazing speed into an almost complete theory — as had the incredibly gifted Cambridge student Paul Dirac working independently in splendid isolation. By 1933, Nobel prizes had been given to Heisenberg and Dirac, and also to Erwin Schrödinger, whose equivalent wave-mechanical formulation only began to appear in January 1926. But Born was left out. More unfair still, Niels Bohr and

Overlooked: while his colleagues quickly won Nobel prizes, Max Born received little recognition.

Heisenberg subsequently gave him little recognition for his role in developing the widely accepted interpretation of quantum mechanics. Finally, in 1954, long after Born had been forced to leave Germany for Britain, he did receive a Nobel prize.

The prize recognized, at long last, Born's discovery of the statistical interpretation of the wave function. He was the first to recognize the profound departure from classical concepts of causality that quantum mechanics implies. In particular, he recognized that although the Schrödinger equation describes a continuous and causal evolution, it nevertheless makes only statistical predictions about observable events. This was before Heisenberg's discovery of the uncertainty principle and Bohr's formulation of complementarity, the essential ingredients in the Copenhagen interpretation. Bohr, Heisenberg and their entourage did not properly acknowledge this fact for a very long time.

In *The End of the Certain World*, Nancy Thorndike Greenspan paints a rich picture of the social, political and intellectual scene in which Born rose to the academic stratosphere from his birth in 1882 into a prosperous Jewish family. He was not a wunderkind. On the contrary, at the age at which the prodigies Dirac, Heisenberg and Wolfgang Pauli had

become legendary scientists, Born had only earned his doctorate in applied mathematics with a thesis on an unexciting topic, and was yet to realize that cutting-edge theoretical physics would be his forte.

As the book recounts, the German physics community took a long time to recognize the talents of this creative and productive man — quite why has been something of a mystery. The intellectual calibre of the cast that populates the book during the early years of Born's career is stunning to a degree that is not adequately recognized or explained, except in the case of Einstein. Born was a modest man of exceptional but not overwhelming intellectual power (unlike the ever-present David Hilbert), and would have had good reason to be intimidated. He was in a similar position to an outstanding Florentine artist whose work remains well known to this day, but who had to make his way in the company of Michelangelo and Leonardo da Vinci. Born compensated for whatever handicap he felt by becoming something of a workaholic, which over time helped him to master an enormous range of physics and produce a prodigious quantity of research papers and splendid texts — from the elementary to his classic treatise on optics.

As the book explains, Born did not have an easy youth psychologically. As a mature man

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

AKG-IMAGES

he was often not sturdy enough, physically and mentally, to handle the load that he imposed on himself. This was compounded by his sometimes shaky marriage to a fragile and not always faithful woman. That he produced so much despite all this is remarkable. The portrait that emerges is of a refined intellectual of the highest ethical standards, unwilling or unable to advocate effectively on his own behalf — even accepting his wife's demand that their children should not attend the Nobel ceremonies. Although he was a refugee from Nazi Germany, he was profoundly disturbed to see his beloved pure physics spawning nuclear weapons in the hands of some of his most talented students: Robert Oppenheimer, Edward Teller, Victor Weisskopf and the spy Klaus Fuchs.

So does this book fill what is a serious gap in the history of twentieth-century physics? In contrast to the other great figures in the quantum revolution, Born's personality has never been described in any depth, and this part of

the gap is filled exceptionally well. As for Born's science, the job is done for the physicist, who can apply well-informed opinions to grade the significance of the people and discoveries described. But other readers would have benefited from an account of Born's work that did a better job of separating the wheat from the chaff.

All the book's readers would have benefited had the editor insisted that we do not need to know the names of all those Alpine hotels the Borns visited, what they ate there, and a barrage of other details. This mass of minutiae often produces a haze that could make it hard for readers to see what is most exceptional and long-lasting among Born's many achievements. Nevertheless, there is no question that any future work on Born will find this book to be an indispensable study of this major figure in one of the most profound transformations in the history of science. ■

Kurt Gottfried is in the Physics Department, Cornell University, Ithaca, New York 14853, USA.

know who Craig was. According to Burkhardt: "Thorpe, supposing Craig was dead, was astonished to learn that Craig was not only alive but in the audience."

The development of ethology has been greatly influenced by the different personalities involved. Nowhere is this shown more clearly than in the relationship between the two key figures in the field, Lorenz and Tinbergen. The characters of these two men could hardly have been more different. Whereas Lorenz was vain, self-centred, an extrovert and a self-styled philosopher, Tinbergen was modest, an introvert and an empiricist. Lorenz characterized himself as a 'farmer', who mainly observed the domesticated birds that he kept around his own house, in contrast to the 'hunter', Tinbergen, who conducted controlled experiments both in the field and in the laboratory. The contrast between the two men became painfully obvious during the Second World War, when Lorenz was a doctor in the German army and Tinbergen was interred in a detention camp for Dutch intellectuals. Burkhardt devotes an entire chapter to Lorenz's conduct during the Nazi regime, and shows that, blinded by ambition, Lorenz did not distance himself from Nazi doctrine. After the war it took some time for the wounds to heal so that the two friends could resume their scientific partnership.

One of Tinbergen's lasting contributions is the identification of the four main problems in animal behaviour: evolution, function, development and causation. Tinbergen has credited British biologist Julian Huxley with identifying three of these as the main problems in biology, to which he merely added development. On reading Burkhardt's account, however, it seems that Tinbergen was being rather generous towards Huxley. "It is from a failure to distinguish between ultimate cause, immediate

Acting on instinct

Patterns of Behavior: Konrad Lorenz, Niko Tinbergen, and the Founding of Ethology

by Richard W. Burkhardt, Jr

University of Chicago Press: 2005. 636 pp.
\$80, £56 (hbk); \$29, £20.50 (pbk)

Johan J. Bolhuis

The Austrian Konrad Lorenz and Dutchman Niko Tinbergen founded ethology, the study of animal behaviour, at the start of the twentieth century. The history of the 'study of instinct', as it was once known, attracts interest from a range of disciplines, and this naturally affects the viewpoint of books on the subject. *Niko's Nature* (Oxford University Press, 2003), for example, a recent biography of Tinbergen, was written by his one-time friend and pupil Hans Kruuk (for a review see *Nature* **427**, 293–294; 2004). In contrast, Richard Burkhardt is a historian, not an ethologist, so his book *Patterns of Behavior* is quite different. Nevertheless, it is not a dry factual biography of a scientific discipline, but a fascinating and often entertaining account of the life and work of some of ethology's key figures. Burkhardt has done a tremendous job, meticulously analysing and describing the rise of ethology. He consulted a multitude of written sources and interviewed many of the important players.

Modern history of science is not only about scientific concepts, Burkhardt explains, but is just as much about the social context of the individual scientists — what he likes to call "ethology's ecologies". The term 'ethology' was introduced by the American William Wheeler, and US biologists active around the end of the nineteenth century might have contributed

much more to the development of ethology if they'd had the resources and the intellectual freedom to pursue their empirical studies and develop new ideas. A particularly poignant example of this is the pioneering biologist Wallace Craig, who greatly influenced Lorenz but struggled to make ends meet for most of his life. When the great British ethologist William Thorpe lectured at Harvard in 1951, he paid tribute to his American colleagues Charles Otis Whitman, Wheeler and Craig. Thorpe was surprised that only one or two members of his large audience seemed to

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Up close and personal: Konrad Lorenz liked to observe birds in their natural habitats.

cause, and mere necessary machinery, that so much of the barren disputes of biology are due," wrote Huxley. It would seem that, with such a sloppy and essentially misguided interpretation, these disputes would not be solved in a hurry. Unfortunately, even Tinbergen's careful analysis of cause and function could not prevent a confusion of concepts that continues to this day.

Some will say that ethology is no longer a scientific discipline in its own right, but that

depends on who you ask: a behavioural ecologist and a cognitive ethologist might give you different answers. Nevertheless, Burkhardt notes that the core ideas of classical ethology dissipated astonishingly rapidly; few contemporary ethologists would use such concepts as 'action-specific energy', for example. This discarding of outmoded ideas would seem natural for any vibrant scientific discipline.

Burkhardt rightly maintains that it was the empirical and theoretical approach introduced

by Lorenz, Tinbergen and their colleagues that made the study of animal behaviour what it is today. In order to study the genomic or neural mechanisms of behaviour, we need to know how behaviour works, and for that an ethological analysis is crucial. This wonderful book shows very clearly how early ethologists made such analysis possible. ■

Johan J. Bolhuis is in the Department of Biology, Utrecht University, Padualaan 14, 3584 CH Utrecht, The Netherlands.

DANCE

Einstein in motion

Constant Speed

Rambert Dance Company

At Sadler's Wells, London, 24–28 May 2005. UK tour begins September 2005.
www.rambert.org.uk

Alison Wright

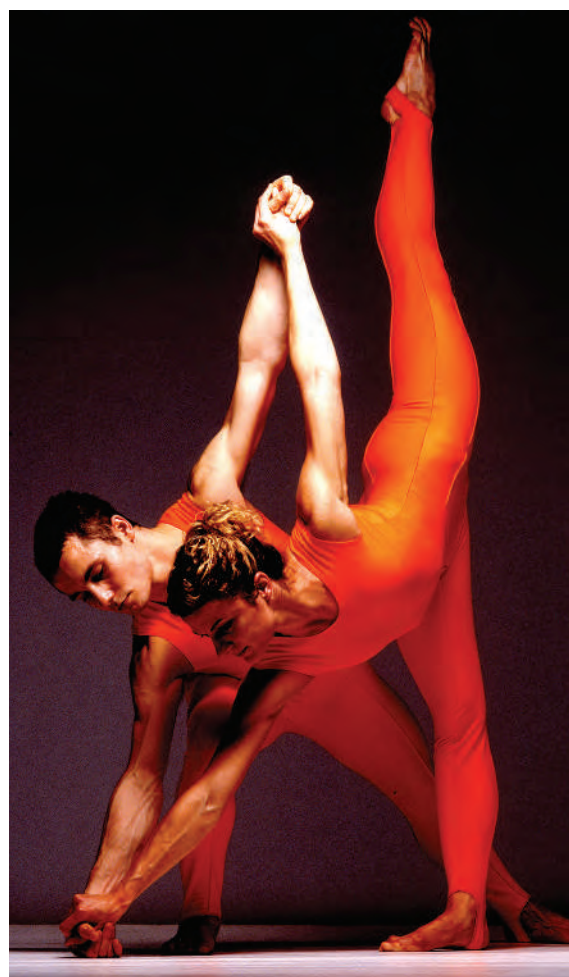
I admit, I cringed when I heard that the UK Institute of Physics (IOP) had commissioned a ballet to celebrate the World Year of Physics 2005, or 'Einstein Year'. Truly enlightening meetings of art and science are rare indeed. *Einstein: The Ballet?* Please, no.

But the IOP had wisely put themselves in the hands of the Rambert Dance Company, and the results, revealed at a Sadler's Wells première on 24 May, are stunning.

Constant Speed is inspired by Einstein's 1905 publications; the World Year of Physics celebrates their centenary. From the patent office in Bern, Switzerland, Einstein dashed off five papers, all of them seminal work, on three themes: brownian motion, the photoelectric effect and the special theory of relativity.

Relativity, and $E = mc^2$, will be forever associated with Einstein. But it is his work on the photoelectric effect, which established the notion of the 'quantum', that Einstein himself regarded as his most revolutionary. The effect describes the release of electrons from a metal when light is shone on its surface. To explain the relation between the energy of the electrons released and the frequency of the incident light, Einstein proposed that light energy is transferred to the electrons in distinct chunks, or quanta. No less significant was his study of fluctuation phenomena within the framework of kinetic-molecular theory — work that recalled the brownian motion seen in the dance of pollen grains in water decades earlier.

Choreographer Mark Baldwin, the artistic director of the Rambert, developed *Constant Speed* through conversations with Ray Rivers, professor of theoretical physics at Imperial College London. Although he claims to be ignorant of physics, Baldwin was struck by a similarity of language — space, time, energy — between physics and dance. Quite rightly, I



Getting physical: dancers of the Rambert Dance Company.

think, he shied away from specifically representing relativity in the piece, alluding only in the title to that theory's central tenet of a constant speed of light. Brownian motion, on the other hand, is a concept easily reflected in the movements and configuration of the dancers. But it is the ideas surrounding the photoelectric effect that dominate *Constant Speed*.

Running on to the stage, fists clenched and elbows pumping, the white-clad dancers could be photoelectrons, released as the seemingly

metallic stage is bathed in light. Or they could be photons, tight packets of energy that also have concerted, wave-like motions. Then the whiteness is subsumed by red, with red lighting and red costumes as more dancers join the fray, evolving eventually through the full spectrum of colour. The colour-coded dancers are compartmentalized — quantized, if you like — each pair exploring its own motion at its own frequency until the rainbow finale. The music by Franz Lehár, composed like Einstein's physics in 1905, adds another dimension: you can almost imagine Einstein day-dreaming, seeing the multi-coloured quanta of his theory before him as he hummed along to the pop music of his day.

At just 27 minutes long, the swirling colour of *Constant Speed* seemed to be over all too quickly. But the work was superbly complemented by the preceding programme, especially *Momenta*, created by Rambert dancer Mikaela Polley. To music by Patrick Nunn, the dance builds steadily in energy, speed and coherence (more physics!). Different moods were created by the tawdry humour of the opening piece, *Judgement of Paris*, and the heart-rending emotion of *Dark*

Elegies. However, the maturity of these pieces, choreographed by Anthony Tudor in the 1930s, perhaps highlighted a few rough edges in *Constant Speed*.

The physics is there if you look — but don't look too hard. As Baldwin said, this is not a physics lesson. Rather, *Constant Speed* is about inspiration: inspiring physics and inspired dance. ■

Alison Wright is the editor of *Nature Physics*.

www.nature.com/naturephysics

A. CRICKMAY

Physics, complexity and causality

Although the laws of physics explain much of the world around us, we still do not have a realistic description of causality in truly complex hierarchical structures.

George F. R. Ellis

The atomic theory of matter and the periodic table of elements allow us to understand the physical nature of material objects, including living beings. Quantum theory illuminates the physical basis of the periodic table and the nature of chemical bonding. Molecular biology shows how complex molecules underlie the development and functioning of living organisms. And neurophysics reveals the functioning of the brain.

In the hierarchy of complexity, each level links to the one above: chemistry links to biochemistry, to cell biology, physiology, psychology, to sociology, economics, and politics. Particle physics is the foundational subject underlying — and so in some sense explaining — all the others. In a reductionist world view, physics is all there is. The cartesian picture of man as a machine seems to be vindicated.

But this view omits important aspects of the world that physics has yet to come to terms with. Our environment is dominated by objects that embody the outcomes of intentional design (buildings, books, computers, teaspoons). Today's physics has nothing to say about the intentionality that has resulted in the existence of such objects, even though this intentionality is clearly causally effective.

A simple statement of fact: there is no physics theory that explains the nature of, or even the existence of, football matches, teapots, or jumbo-jet aircraft. The human mind is physically based, but there is no hope whatever of predicting the behaviour it controls from the underlying physical laws. Even if we had a satisfactory fundamental physics 'theory of everything', this situation would remain unchanged: physics would still fail to explain the outcomes of human purpose, and so would provide an incomplete description of the real world around us.

Can we nevertheless claim that the underlying physics uniquely causally determines what happens, even if we cannot predict the outcome? To examine whether we can, contemplate what is required for this claim to be true within its proper cosmic context. The implication is

that the particles existing when the cosmic background radiation was decoupling from matter, in the early Universe, were placed precisely so as to make it inevitable that 14 billion years later, human beings would exist, Charles Townes would conceive of the laser, and Edward Witten would develop string theory. Is it plausible that quantum fluctuations in the inflationary era in the very



Intelligent design: no physics theory is able to explain a teapot.

early Universe — the source of the perturbations at the time of decoupling — implied the future inevitability of the Mona Lisa and Einstein's theory of relativity? Those fluctuations are supposed to have been random, which by definition means without purpose or meaning.

However, such meaning did indeed come into being. Ever higher levels of interaction and causality arose as complexity spontaneously increased in the expanding Universe, allowing life to emerge. Darwinian processes of selection guided the physical development of living systems, including the human brain.

It is possible that what actually happened was the contextual emergence of complexity: the existence of human beings and their creations was not uniquely implied by the initial data in the early Universe; rather the underlying physics together with that initial data created a context that made the existence of human beings possible. Conditions at the time of the decoupling of matter and radiation 14 billion years ago were such as to lead to the eventual development of minds that are autonomously effective. Such minds are able to create higher-level order, such as the Hubble Space Telescope and Kurt Gödel's incompleteness theorem, that embodies a purpose and meaning not in existence before.

With this view, the higher levels in the hierarchy of complexity have autonomous causal powers that are functionally independent of lower-level processes. Top-down causation takes place as well as bottom-up action, with higher-level contexts determining the outcome of lower-level functioning, and even modifying the nature of lower-level constituents.

Stored information plays a key role, resulting in non-linear dynamics that are non-local in space and time. Brain functioning is causally affected by abstractions, such as the value of money, the rules of chess and the theory of the laser. These abstractions are realized as brain states in individuals, but are not equivalent to them — James Clerk Maxwell's theory of electromagnetism is not the same as any individual's brain state. Although such concepts are causally effective, they are not themselves physical variables. Consequently physics *per se* cannot causally determine the outcome of human creativity; rather it creates the 'possibility space' to allow human intelligence to function autonomously.

This situation is not dependent specifically on human intentionality. Physics by itself cannot explain any behaviour that is adaptive and depends on context, for example, beaver dam-building and the dances of bees. It is plausible that these also emerged at late times in the expanding Universe as higher-level autonomous behaviours, made possible but not causally determined by the underlying physics and chemistry of matter.

If this is the case, the challenge to physics is to develop a realistic description of causality in truly complex hierarchical structures, where top-down causation and memory effects allow autonomous higher levels of order to emerge with genuine causal powers. So far, attempts to relate physics to complexity — such as the reaction-diffusion equation, chaos theory, the renormalization group, complexity theory — take us only a small step on this road. ■

George F. R. Ellis is at the Mathematics Department, University of Cape Town, Rondebosch, Cape Town 7701, South Africa.

FURTHER READING

Ellis, G. F. R. *Phys. Today* (in the press).
Bishop, R. C. *Phil. Sci.* (in the press).

CANCER GENOMICS

Small RNAs with big impacts

Paul S. Meltzer

Although they are tiny, microRNAs can have large-scale effects because they regulate a variety of genes. These minuscule molecules are now definitively linked to the development of cancer.

During the past few years, molecular biologists have been stunned by the discovery of hundreds of genes that encode small RNA molecules¹. These microRNAs (miRNAs) — 21 to 25 nucleotides in length — are negative regulators of gene expression. The mechanisms by which they work are similar in plants and animals, implying that they are involved in fundamental cellular processes². As cancer is essentially a consequence of disordered genome function, one might expect these regulatory molecules to be involved in the development of this disease. Indeed, there are hints that the levels of some miRNAs are altered in cancer^{3,4}; there is also evidence that an miRNA regulates the cancer-promoting *ras* gene⁵. Three studies in this issue^{6–8} change the landscape of cancer genetics by establishing the specific miRNAs expressed in most common cancers, and investigating the effects of miRNAs on cancer development and cancer genes.

The initial product of an miRNA gene (Fig. 1) goes through several processing steps before it is exported from the nucleus to the cytoplasm. One strand of the resulting double-stranded RNA is then incorporated into the 'RNA-induced silencing complex' (RISC). RISC can target protein-coding messenger RNAs (mRNAs) either for inhibition, by blocking their translation into protein, or destruction (as in RNA interference). Base pairing between the miRNA and its complementary target mRNA gives the process its specificity.

The choice between translational inhibition and destruction is thought to be governed by the degree of mismatch between the miRNA and its target mRNA, with degradation being the outcome for best-matched targets. Because miRNAs can inhibit the translation of imperfectly matched targets, it is possible that each miRNA may target multiple genes, and that several miRNAs may regulate a given target. The interplay between mRNA and miRNA is vital for the normal regulation of gene expression, but its role in disease is just beginning to be investigated.

The expression of miRNAs appears to be highly regulated according to the cell's

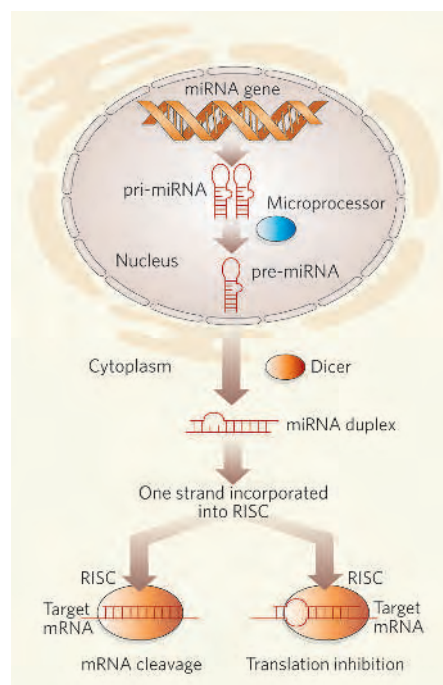


Figure 1 | MicroRNA production. The precursor of an miRNA (pri-miRNA) is transcribed in the nucleus. It forms a stem-loop structure that is processed to form another precursor (pre-miRNA) before being exported to the cytoplasm. Further processing by the Dicer protein creates the mature miRNA, one strand of which is incorporated into the RNA-induced silencing complex (RISC). Base pairing between the miRNA and its target directs RISC to either destroy the mRNA or impede its translation into protein. The initial stem-loop configuration of the primary transcript provides structural clues that have been used to guide searches of genomic sequence for candidate miRNA genes.

developmental lineage and stage. Lu *et al.* (page 834)⁶ find that aspects of this specificity are maintained in cancer: by measuring the expression of 217 human miRNAs in cancer samples, they found that the pattern of miRNA expression varies dramatically across tumour types. Remarkably, the expression pattern of this small set of miRNAs defines the cancer type better than expression data from 16,000 mRNAs. As might be expected from the role of

some miRNAs in development, the miRNA profiles of tumours are in accord with the tumours' developmental history; tumours derived from tissues with a common embryonic precursor (such as gastric, colon and liver cancers, which are all derived from the embryonic endoderm) share similar miRNA expression patterns. Leukaemias are clearly separate from solid tumours and, strikingly, are subgrouped according to their underlying genetic abnormalities.

These observations could improve the diagnosis of poorly defined cancers with unknown origins, allowing better-informed choices for treatment. They also promise to shed light on the regulatory circuits that malfunction during tumorigenesis. The expression of miRNAs seems to be lower in cancers than in normal tissues — consistent with the possibility that reduced miRNA expression leads to a cancer-specific block that halts the normal development of cells. This may allow the cells to continue to divide and grow, unlike their mature counterparts. Small RNAs can easily be measured from the formalin-fixed tissue specimens used routinely in hospital pathology laboratories; so potential miRNA-based diagnostics could fit simply into the standard hospital workflow.

He *et al.* (page 828)⁷ examine an old puzzle: the accumulation of extra copies of a chromosome 13 fragment (13q31–32) in certain human lymphomas (cancers derived from immune cells). The region of amplification contains a gene called *c13orf25* (ref. 9), which turns out to encode the precursors of seven miRNAs. He *et al.* establish that the *c13orf25* miRNAs are indeed overexpressed in lymphoma cells that have extra copies of *c13orf25*.

To investigate the biological consequences of this miRNA overexpression, they employ a genetically engineered mouse model in which lymphoma development is driven by a cancer-promoting gene (oncogene) called *myc*. The authors infected blood-forming cells from these mice with a retrovirus carrying a portion of the miRNA cluster and then transplanted the cells back into recipient lymphoma-prone mice. Compared with control mice injected with a virus lacking the miRNAs, there was a

striking decrease in the latency of leukaemia development (from 3–6 months to 51 days) and an increase in the frequency of the cancer in the mice (from about 30% to 100% of animals). The mechanism for these effects remains unclear, and it is uncertain which components of the c13orf25 miRNA cluster are responsible. Nonetheless, this study emphatically nominates this miRNA cluster as the first candidate non-coding oncogene.

The connection between miRNAs and the *myc* oncogene is also examined by O'Donnell *et al.* (page 839)⁸. This gene encodes the transcription regulator c-Myc, the overexpression of which is frequent in cancer. To explore the effects of c-Myc on miRNAs, O'Donnell *et al.* use a lymphoma cell line carrying a *myc* gene that can be switched on by treating the cells with an 'inducer'. They find that increasing the expression of c-Myc leads to increased expression of six miRNAs. Remarkably, two of these are encoded by the c13orf25 cluster and the remainder are encoded by two related clusters on chromosomes 7 and X. O'Donnell *et al.* confirm that c-Myc binds to a candidate regulatory site in c13orf25.

A predicted target of two of the encoded miRNAs is the transcription factor E2F1, which is itself a critical regulator of the cell cycle. O'Donnell *et al.* demonstrate that this prediction is correct, and that expression of E2F1 is only affected if the miRNA target site is present in the E2F1 mRNA. E2F1 and c-Myc are known to induce each other's expression. In the absence of other controls, this could set up a positive-feedback loop leading to overexpression of both genes — with disastrous consequences for normal cell-cycle regulation. O'Donnell *et al.* propose that by negatively regulating E2F1, miRNAs induced by c-Myc could dampen the runaway effect, fine-tuning the dynamics of E2F1 action during the cell cycle.

Questions remain, of course. What regulates the expression of miRNAs? What are the targets of each miRNA? Do miRNAs act mainly to 'fine-tune' gene expression or more often as binary on/off switches? Given that each miRNA may regulate numerous targets, it is possible that thousands of protein-coding genes could be regulated by a few hundred miRNAs. Candidate miRNA target genes can be identified by bioinformatics approaches, but a great deal of experimental work remains to be done in validation.

We need to find out which of the biological pathways underlying cancer are regulated by miRNAs, but the complexity of the problem is underlined by the dual roles for the same miRNAs identified in two of this week's reports^{7,8}. In one context, the c13orf25 cluster can act as an oncogene; in another it seems to antagonize the effects of different oncogenes, acting like a classic tumour-suppressor gene. Sorting out the miRNA regulatory networks will be challenging, but is vital to explain the nuanced regulation of gene expression

essential to the growth, development and survival of multicellular organisms.

Paul S. Meltzer is at the Cancer Genetics Branch, National Human Genome Research Institute, 50 South Drive, Bethesda, Maryland 20892-8000, USA.
e-mail: pmeltzer@nhgri.nih.gov

1. Lagos-Quintana, M., Rauhut, R., Lendeckel, W. & Tuschl, T. *Science* **294**, 853–858 (2001).

2. Bartel, D. P. *Cell* **116**, 281–297 (2004).
3. Calin, G. A. *et al.* *Proc. Natl Acad. Sci. USA* **101**, 11755–11760 (2004).
4. Eis, P. S. *et al.* *Proc. Natl Acad. Sci. USA* **102**, 3627–3632 (2005).
5. Johnson, S. M. *et al.* *Cell* **120**, 635–647 (2005).
6. Lu, J. *et al.* *Nature* **435**, 834–838 (2005).
7. He, L. *et al.* *Nature* **435**, 828–833 (2005).
8. O'Donnell, K. A., Wentzel, E. A., Zeller, K. I., Dang, C. V. & Mendell, J. T. *Nature* **435**, 839–843 (2005).
9. Ota, A. *et al.* *Cancer Res.* **64**, 3087–3095 (2004).

ORGANIC CHEMISTRY

Fast reactions 'on water'

Jaap E. Klijn and Jan B. F. N. Engberts

Efficient reactions in aqueous organic chemistry do not require soluble reactants, as had been thought. A newly developed 'on-water' protocol is characterized by short reaction times, and the products are easy to isolate.

Water is unique. The chemistry of living organisms depends on its combination of unusual properties, and it is difficult to imagine life in the absence of the aqueous medium¹. For synthetic organic chemistry, however, it is less important, because water is traditionally not a popular choice of solvent.

There are two main reasons for this. Functional groups in the organic molecule may themselves react with water and, more importantly, most organic molecules are nonpolar, and so hydrophobic and generally highly insoluble in water. It is therefore assumed that a mixture of water and one or more non-polar organic reactants will usually have low reaction rates and low yields of the desired products. As a group led by Barry Sharpless reports in *Angewandte Chemie*², however, this assumption does not hold for a variety of organic reactions in the 'on water' approach that they have pioneered.

There was already compelling evidence that many organic reactions are faster in water than in organic solvents, but that evidence was based on reaction kinetics in very dilute, homogeneous solutions (in which the

reactants are in the same phase). Breslow and Rideout³ were the first to show that the otherwise solvent-insensitive Diels–Alder reactions — which are among the most useful reactions in organic chemistry, often used for the synthesis of six-membered rings — may be greatly accelerated in water. Similar results were obtained for other types of simple (uni- and bimolecular) organic reactions. Subsequent mechanistic studies have established that this behaviour results from enforced hydrophobic interactions and stabilization of the activated complex by hydrogen-bond formation. Although synthetic applications evolved from these studies⁴, there were invariably limitations that stemmed from lack of solubility. And scant attention was paid to the kinetics of organic transformations under heterogeneous aqueous conditions, when the reactants are in different phases.

Sharpless's group² now shows that several uni- and bimolecular reactions are greatly accelerated when carried out in vigorously stirred aqueous suspensions (Fig. 1). The reactions include the important classes known as cycloadditions, ene reactions,

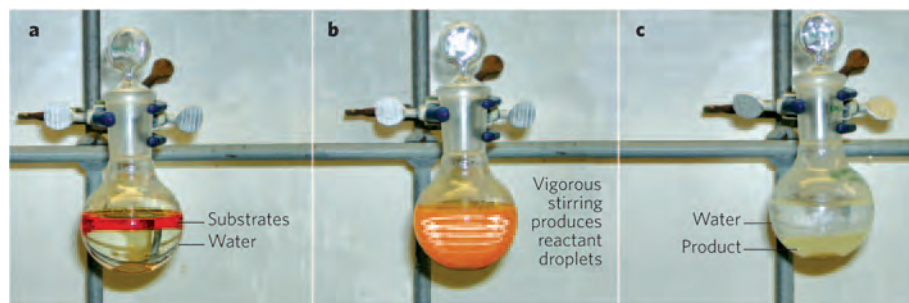


Figure 1 | The 'on water' protocol, as developed by Sharpless and colleagues². **a**, The organic substrates are initially floating on top of the water. **b**, The mixture is stirred vigorously, which disperses the reactants as small droplets and leads to a large increase in the surface area between the reactants and the aqueous phase. **c**, The reaction product either precipitates as a solid or floats on top of the water as an oil, either of which can be isolated easily. The mechanism of the rate acceleration has yet to be clarified, but it probably depends on the increase in interfacial area.

Claisen rearrangements and nucleophilic substitutions. The reactants are initially floating on the surface of water, hence the 'on water' designation, and the reaction product is often produced in a pure state under or on the water, allowing it to be easily isolated by phase separation or filtration. This simple protocol can be carried out in a beaker, as well as in more complex vessels.

Sharpless and colleagues' experiments were performed with one or two liquid, water-insoluble reaction partners or, occasionally, a mixture of one liquid and one solid. Although the authors have not undertaken detailed kinetic experiments, the yields of pure products after varying reaction times convincingly demonstrate that the rates are higher than those under solvent-free ('neat') or homogeneous conditions. One reaction — the $[2\sigma+2\sigma+2\pi]$ cycloaddition of quadricyclane with dimethyl azodicarboxylate (DMAD) to yield a 1,2-diazetidene as the single product — provides, from a scientific point of view, a good example (Fig. 2). The time to completion for the 'on-water' reaction is only 10 minutes at 23 °C; that compares with 48 hours for the solvent-free reaction, and 18 hours in methanol and more than 120 hours in toluene.

The efficiency of 'on-water' reactions is amazing, but the mechanism remains unclear. In that respect, some of Sharpless and colleagues' observations may help. First, 'on-water' reactions of quadricyclane with diethyl azodicarboxylate that are carried out in the additional presence of apolar solvents, such as toluene, remain much faster than the same reactions in toluene alone.

Second, there is further evidence that the heterogeneous reaction is more efficient. Gradual addition of methanol to the aqueous phase had little effect. If, at a critical methanol concentration, the heterogeneous system switches to a homogeneous one, however, a dramatic reduction in reaction rate results. But the replacement of water by C_6F_{14} , so providing 'on perfluorohexane' heterogeneous conditions, produced no significant rate enhancement for the reaction of quadricyclane with DMAD. This experiment shows that water is required, and that heterogeneity is not in itself the secret of 'on-water' efficiency.

Further mechanistic studies should provide more evidence about the exact origin of the rate enhancements. The Sharpless group also carried out the reaction of quadricyclane with DMAD using deuterium oxide (D_2O), and found that the time to completion was longer (45 minutes, not 10 minutes, as with H_2O). This is a challenging observation, because no rate-determining proton, hydrogen or hydride transfer is involved. Perhaps the difference in viscosity between water and deuterium oxide is involved here. Assuming that the reactions occur at the surface of the reactant particles in suspension, rather than in the aqueous phase at low reactant concentrations, particular emphasis should be placed on

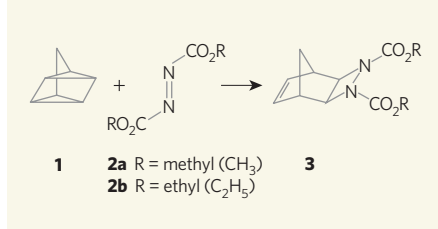


Figure 2 | Rapid reaction. The reaction of quadricyclane (1) with dimethyl (2a) or diethyl (2b) azodicarboxylate to yield 1,2-diazetidene (3). With the 'on-water' protocol, the reaction runs to completion in only 10 minutes, much faster than with other procedures.

investigating the properties of water and the reactants at these interfaces. Because most of the activated complexes in the reactions examined by Sharpless and colleagues are probably more efficient at forming hydrogen bonds than are the initial states, the hydration layers of the particles may play a crucial role.

The scope of these new findings² will have to be extended to take the 'on-water'

procedure to the forefront of organic chemistry. One challenge will be to devise a protocol that allows the successful use of two solid reactants. But even at this stage the approach will have an impact on synthetic organic chemistry in the laboratory, and possibly on larger scales, and will also influence thinking about chemical reactivity in heterogeneous aqueous media. It brings biochemistry and organic chemistry closer together in the beneficial use of water as the reaction medium.

Jaap E. Klijn and Jan B. F. N. Engberts are in the Physical Organic Chemistry Unit, Stratingh Institute, University of Groningen, 9747 AG Groningen, The Netherlands. e-mails: Jaapklijn@home.nl; J.B.F.N.Engberts@rug.nl

1. Daniel, R. M., Finney, J. L. & Stoneham, M. (eds) *Phil. Trans. R. Soc. Lond. B* **359**, 1141–1328 (2004).
2. Narayan, S. et al. *Angew. Chem. Int. Edn* **44**, 3275–3279 (2005).
3. Breslow, R. & Rideout, D. C. *J. Am. Chem. Soc.* **102**, 7816–7817 (1980).
4. Grieco, P. A. (ed.) *Organic Synthesis in Water* (Blackie, London, 1998).

STRUCTURAL BIOLOGY

Prying into prions

Christopher M. Dobson

Various aberrant protein forms are the subject of intense research. It is not easy to probe their structures, but studies that have done so provide telling information about their biological properties.

One of the most intriguing issues in biology is the occasional conversion of proteins from their intricately folded functional forms into thread-like molecular aggregates. These transformations into an alternative form of protein structure¹ are of much more than academic interest — such aggregates are linked to some of the most feared diseases of the modern era², and to the previously heretical idea that transmission of genetic information can occur without the involvement of nucleic acids³. Three reports in this issue^{4–6} provide provocative data that illuminate the structures of these types of aggregate, and suggest how such structures might explain their extraordinary properties.

The alternatives to the familiar forms of proteins are usually known as amyloid (or amyloid-like) fibrils and are perhaps most notorious for their association with Alzheimer's disease^{1,7}. But they are also involved in some 20 other 'misfolding' disorders⁷. These include type II diabetes, as well as the transmissible forms of the diseases epitomized by scrapie and 'mad cow' disease in domesticated animals, and by kuru and Creutzfeldt–Jakob disease in humans^{2,8}. The proteins involved in these conditions are known as prions (proteinaceous infectious

particles). Prions are increasingly turning up in different organisms, particularly yeast and other fungi. The yeast prions are not functionally or structurally related to their mammalian namesakes, and their ability to convert into fibrillar aggregates is coupled not just to disease but also to the inheritance of genetic traits^{3,8}.

Proteins in amyloid fibrils are folded to produce a core region consisting of a continuous array of β -sheets. Such sheets are a familiar type of protein motif, and here are made up of β -strands that are oriented perpendicular to the fibril axis in an arrangement called a cross- β structure⁹. The ability to form this type of structure may be a generic feature of polypeptide chains¹, although the specific amino-acid sequence of the chain affects both the propensity to form fibrils and the way a given molecule is arranged within the fibrils. Knowledge of this latter aspect is vital for understanding the properties of protein forms such as prions, but has been seriously limited by the intractability of amyloid fibrils to the traditional methods of structural biology.

Nelson et al. (page 773)⁴ now report a breakthrough in this context. They managed to persuade a seven-residue peptide from one end — the amino-terminal region — of a yeast prion known as Sup35 to form three-



50 YEARS AGO

"The English Climate" — In this engagingly written volume Dr. C. E. P. Brooks describes the climate of Britain in relation to the major factors controlling it as well as the various weather processes and seasonal vicissitudes which make it up... In the chapter on "Fog and Soot" much prominence is given to the disastrous 'smog' of December 1952. The heavy death-rate from bronchitis and pneumonia is apparently attributed in the main to sulphur dioxide and soot... No reference is made in this chapter to modern smoke-abatement practices and changed methods of domestic heating as bearing on the experience of the older generation that London fogs to-day have lost the sooty blackness of Victorian times.

From *Nature* 11 June 1955.

100 YEARS AGO

An article entitled "Some Candid Impressions of England" is contributed to the current number of the *National Review* by a "German Resident". The first fact which strikes the contributor is the indifference of Englishmen to their duties as citizens of a great Empire, and it seems to him, looking at English schools, that the mainspring of German success is here. He says:—"Our youths, like your youths, are human, and would be lazy if there were no penalty for idleness... I look at England and see the want of such an influence even in your public schools, which are good in a way, so far as they form character, but bad in that they neglect intellect."... The majority of our workers, he remarks, read little but the sporting Press, and care for little but betting and sport. It is pointed out that the Germans have destroyed in this generation the superstition that Germany makes only poor and cheap articles. "Our Mercedes motors and scientific and optical instruments are the best and most expensive in the world, and no English article of their class can for a moment compete with them."

From *Nature* 8 June 1905.

dimensional microcrystals that eventually proved to be suitable for high-resolution X-ray diffraction studies.

In the microcrystals, the peptide molecules are assembled into a structure that, remarkably, seems to possess the key features of the core of an amyloid fibril. Indeed, observations with electron microscopy suggest that the crystals result from the assembly of nanometre-sized components. The cross- β structure seen in the crystals is built up of pairs of β -sheets, so that side chains of amino-acid residues on the inner side of each pair of sheets are enmeshed with those of the residues of the other sheet to such a degree that water is excluded (Fig. 1). The outer faces of the pairs of sheets are hydrated and more distant from each other, implying that this less intimate interaction could be a crystal contact rather than a feature of the fibrillar state. The suggestion that the fundamental structural unit of a fibril could be just a pair of rather flat sheets is consistent with conclusions from studies of amyloid fibrils of several proteins, including insulin, by cryo-electron microscopy¹⁰ — so adding further support to the supposition that the inherent fibrillar interactions have been captured in the crystalline molecular array.

One of the main objectives in prion research is to establish the relationship between specific structural features and biological properties such as infectivity. Ritter *et al.* (page 844)⁵ set out to address this question using the 71-residue carboxy-terminal region of a prion, known as HET-s, from a filamentous fungus. This protein could not be crystallized, but, by using nuclear magnetic resonance techniques, four regions of the sequence were identified as being involved in β -sheet structure. Each region consists of 7–10 contiguous residues and, by judicious use of fluorescent probes, the amyloid core structure was again found to be a pair of β -sheets.

Following model-building and sequence analysis, Ritter *et al.* came up with a proposal for the arrangement of the polypeptide chain within the fibril core that shows marked similarities to the way the seven-residue peptide is arranged in the crystals studied by Nelson *et al.*⁴. Finally, by making a series of mutations designed to disrupt regions of the protein in turn, the amyloid core structure was indeed shown to be the key element required to maintain infectivity in living cells.

Further secrets about the way in which the biological properties of prions are linked to structure are revealed by Krishnan and Lindquist (page 765)⁶. They studied a much larger (250-residue) fragment of Sup35 that has all the characteristics of the natural yeast prion. Like Ritter *et al.*⁵, in the absence of crystals they used a range of clever methods to probe the environment of individual residues; in particular, they used a fluorescence technique that enabled the relative proximities of different residues in the sequence to be defined. By analysing the large array of data

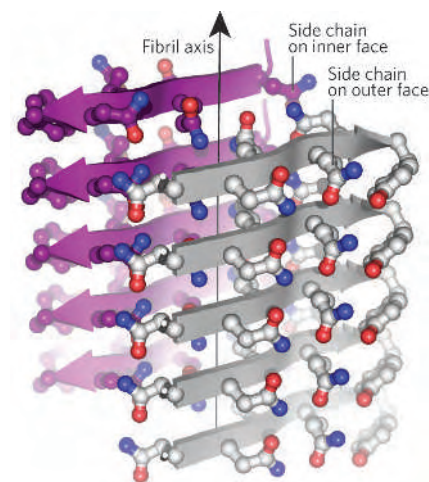


Figure 1 | Structure of the seven-residue peptide from the yeast prion, Sup35. The figure shows the double β -sheet arrangement observed in the three-dimensional crystal, which could represent the stacking of peptide molecules in an amyloid fibril. Here, the individual molecules form the β -strands (purple and grey arrows) that lie perpendicular to the fibrillar axis and are linked by hydrogen bonds to form a pair of β -sheets; addition of peptides to the edges of the sheets elongates the structure in the direction of the arrow and generates a fibril. On the inner face of each sheet, the side chains interact to such an extent that water is excluded. On the outer face, the side chains are hydrated in the crystal and would be exposed to solvent in the fibril. Longer polypeptides can form similar structures by folding back on themselves, with parts of the sequence being in the core structure and the rest forming turns and loops¹⁰. (Modified from Fig. 2a of ref. 4.)

collected in this heroic study, Krishnan and Lindquist identified two regions of the sequence, each some 15 residues long, that form the key intermolecular interactions in the fibrillar state of the prion. They were then able to show that the first step in the aggregation process involves the formation of a collapsed and disorganized 'molten' structure, in which nucleation can occur and allow the development of the fibrillar structure.

Krishnan and Lindquist's study⁶ starts to get to the heart of the unique properties of prions, as it probes the mechanism that triggers their ability to self-propagate. One other aspect of prions that has stimulated much debate is the existence of 'strains', such that a given prion can be linked with a variety of self-perpetuating traits or forms of disease^{8,11}. In this context, Krishnan and Lindquist made a striking discovery — they found that the specific residues making up the core of the fibrils are different when the fibrils are grown under different conditions, here simply a change in temperature of about 20 °C. It has already been shown that fibrils with distinctive characteristics can be seeded and propagated in much the same manner as crystals¹². This new observation therefore suggests that different strains of prions could be linked to specific and self-

propagating variations in the molecular interactions within the highly organized fibrils. And this conclusion hints that the crucial barriers that limit the transmission of prion diseases between different mammalian species^{2,8} might have similar molecular explanations.

Much remains to be learned about the nature of prions, and indeed about the more general and increasingly pervasive amyloid phenomenon. But these three studies^{4–6} show yet again how the power of hard-won structural knowledge stimulates the elucidation of even the most complex of biological phenomena.

Christopher M. Dobson is in the Department of

Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK.
e-mail: cmd44@cam.ac.uk

1. Dobson, C. M. *Nature* **426**, 884–890 (2003).
2. Prusiner, S. *Science* **278**, 245–251 (1997).
3. Tuite, M. F. & Cox, B. S. *Nature Rev. Mol. Cell Biol.* **4**, 878–890 (2003).
4. Nelson, R. *et al.* *Nature* **435**, 773–778 (2005).
5. Ritter, C. *et al.* *Nature* **435**, 844–848 (2005).
6. Krishnan, R. & Lindquist, S. L. *Nature* **435**, 765–772 (2005).
7. Selkoe, D. J. *Nature* **426**, 900–904 (2003).
8. Caughey, B. *Nature Med.* **6**, 751–754 (2000).
9. Sunde, M. & Blake, C. C. F. *Adv. Protein Chem.* **50**, 123–159 (1997).
10. Jimenez, J. L. *et al.* *Proc. Natl Acad. Sci. USA* **99**, 9196–9201 (2002).
11. Tanaka, M., Chien, P., Naber, N., Cooke, R. & Weissman, J. S. *Nature* **428**, 323–328 (2004).
12. Petkova, A. T. *et al.* *Science* **307**, 262–265 (2005).

PLANETARY SCIENCE

Shades of Titan

Louise Prockter

Instruments aboard the Cassini spacecraft can ‘see’ through the dense atmosphere of Titan, Saturn’s largest moon. Returned images hint at the existence of features such as ridges and valleys, and perhaps an icy volcano.

Wreathed in clouds and haze, Titan has long been the focus of Earth-based telescopes, and their probings have yielded tantalizing clues about the nature of the moon’s surface. The main obstacle to studying Titan’s surface is its atmosphere of nitrogen and methane — peeking through the dense fog is a tough task. The Huygens probe has provided the first close-up images of Titan’s landscape¹, but, flying above the atmosphere, the Cassini spacecraft also carries instruments for observing the surface. On page 786 of this issue², Sotin and collaborators reveal images acquired at visible to near-infrared wavelengths that open a new window onto Titan’s surface.

A quarter of a century after the two Voyager spacecraft visited the saturnian system, the Cassini–Huygens spacecraft — a joint mission by NASA and the European and Italian space agencies — has returned for a closer look at the giant planet, its rings, space environment and satellites, especially its largest moon Titan. The Cassini orbiter has a full arsenal of instruments designed to do the equivalent of sniffing, seeing and tasting from above the atmosphere. One such instrument is the sophisticated Visible and Infrared Mapping Spectrometer (VIMS), whose primary objective is to determine the composition and distribution of materials on the surfaces of Saturn’s icy satellites. VIMS measures reflected and emitted radiation using filters at wavelengths in the visible to mid-infrared range. At some of those wavelengths, the methane atmosphere of Titan is effectively transparent, allowing light reflected from the surface to reach the detector (Fig. 1).

So far, Cassini has made five close flybys of Titan. Along with images returned by the Huygens probe during its descent through the atmosphere¹ to the surface, the Cassini data show that Titan’s landscape is a mélange of bright and dark material, possibly sculpted by processes such as wind, tectonic movement, fluid flow, and asteroid or meteorite impacts³.

Sotin *et al.*² analysed a region of Titan’s surface some 150 km × 150 km in area, using VIMS images that were of sufficient resolution to discriminate features a few kilometres in scale. The authors compared two or more

images of the same area acquired through different filters to produce image ratios that emphasize variations in surface composition. In some regions of the spectrum, the bright material observed is nearly twice as bright as the dark material, but both are spectrally similar — meaning that they may actually be of similar composition. Moreover, Sotin and colleagues’ analysis suggests that, even in bright regions, there is no water ice exposed at the surface.

By comparing brightness variations at different wavelengths with some features observed in the image ratios, the authors conclude that some of the surface heterogeneity may be due to the presence of slopes rather than material of different composition. They derive crude topography from the images, and propose that ridges and valleys are present with relief on the order of a few hundred metres and slopes of about 10°. If correct, this interpretation might explain the absence of marked compositional variations between the bright and dark material in this area, and may also rule out the proposed existence of pools of liquid methane larger than a kilometre (the scale of these images).

Perhaps the most striking feature in the VIMS images is a bright, circular region, 30 km in diameter, with what seem to be two horn-shaped markings pointing to the west (Fig. 1, inset). Sotin and collaborators attribute this feature to an icy volcanic dome, formed by an upwelling plume of ‘hot ice’ (in this case perhaps not water ice, but something more exotic, such as nitrogen ice that has broken through the surface). Although this is an intriguing hypothesis, the images are not of sufficient resolution to provide details below a few hundred metres, and the feature may well turn out to be an impact crater. On other worlds where icy volcanic resurfacing is thought to have occurred, such as on Jupiter’s moon Ganymede, high-resolution imaging

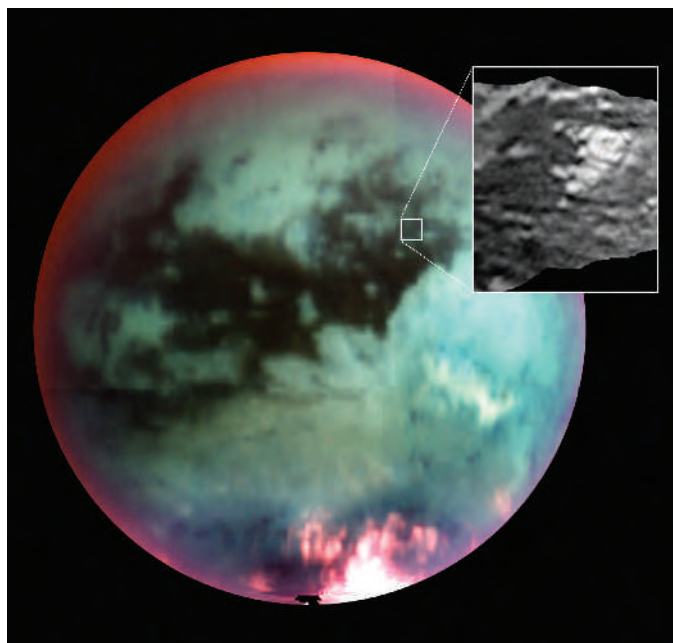


Figure 1 | Titan uncloaked. This false-colour mosaic of Titan was constructed using six medium-resolution (25 km per pixel) infrared images obtained during the Cassini flyby of 26 October 2004. The colours correspond to atmospheric (red) and surface (green and blue) features that are not visible to the human eye. The inset shows a high-resolution (2 km per pixel) image taken using a 2.30- μ m filter near the point of Cassini’s closest approach to Titan (1,200 km). Sotin *et al.*² interpret the prominent circular feature as an icy volcano with lobate flows extending westwards.

has yielded ambiguous evidence for putative cryovolcanic structures, and no direct indication of any source vents⁴. If volcanism were indeed responsible for this circular feature on Titan, this would have significant implications for the moon's composition and for the heat sources necessary to form such features from icy lavas.

Cassini carries another instrument, the Titan Radar Mapper, which can penetrate the clouds at different wavelengths from VIMS but at similar resolutions. The Radar Mapper has sent back images of lobate and heterogeneous features, and in another paper⁵ Elachi *et al.* interpret these features as icy flows, sinuous channels and volcanic constructs. The radar instrument can also measure some properties of Titan's surface — early results⁵ indicate that the surface consists of porous ice or organic sludge, while radar-dark patches may represent frozen hydrocarbons.

Observations made with VIMS and the Radar Mapper, along with data from Cassini's

Imaging Science Subsystem³, are complementary and will reveal much about Titan's evolution and the interactions between its surface, atmosphere and the saturnian environment. Whatever the bright and dark regions on Titan's surface turn out to be, Cassini is taking us into a new and thrilling era of satellite exploration. With 40 more planned close flybys of Titan, as well as of several other saturnian moons, we are only at the beginning of this fantastic journey. ■

Louise Prockter is in the Planetary Exploration Group, Applied Physics Laboratory, Johns Hopkins University, 11100 Johns Hopkins Road, Laurel, Maryland 20723, USA.
e-mail: Louise.Prockter@jhuapl.edu

1. Cassini-Huygens mission *ESA Bull.* **121**, 6–9 (2005).
2. Sotin, C. *et al.* *Nature* **435**, 786–789 (2005).
3. Porco, C. C. *et al.* *Nature* **434**, 159–168 (2005).
4. Pappalardo, R. T. *et al.* in *Jupiter: The Planet, Satellites and Magnetosphere* (eds Bagenel, F., Dowling, T. E. & McKinnon, W. B.) 363–397 (Cambridge Univ. Press, 2004).
5. Elachi, C. *et al.* *Science* **308**, 970–974 (2005).

PLANT-FUNGAL ASSOCIATIONS

Cue for the branching connection

Martin Parniske

Plant roots release potent molecules that activate symbiotic fungi and initiate a harmonious relationship. It turns out that the same compounds are detected by parasitic weeds for less benign purposes.

Of all the fungi, it is a group known as the Glomeromycota that arguably has the greatest claim to public notice. The reason is that its members form symbiotic associations known as arbuscular mycorrhiza with plant roots, and without this symbiosis our planet's flora would have a very different constitution. The latest episode in understanding how plants communicate with their fungal partners appears on page 824 of this issue¹. There, Akiyama *et al.* describe the characterization of a factor that is released by the plant as a component of the symbiotic dialogue.

It seems likely that arbuscular mycorrhiza had a key role in shaping life on Earth by helping the first primitive plants to colonize land^{2,3}. Plant biomass constitutes the base of food chains, and was the prerequisite for the evolution of land-based animals and ultimately for our own existence. Early land plants did not have elaborate root systems for extracting nutrients from the soil, and therefore probably relied on fungal symbioses to explore the substrate and deliver those nutrients. In return, the fungi obtained carbohydrates produced by photosynthesis. Because arbuscular mycorrhiza are so beneficial, all major plant lineages have retained the association — more than 80% of higher plant species form a symbiotic relationship with these fungi, which are

found in most terrestrial ecosystems⁴.

Probably as a consequence of millions of years of coevolution, arbuscular mycorrhizal fungi rely so much on the interaction with plants that they cannot be grown in isolation. A spore buried in the ground can only ramify through the branching threads known as hyphae if it encounters a host root. The spores will germinate on agar surfaces in Petri dishes. In the absence of a host root, however, the germ tubes become inactive and their cytoplasm retracts into the spore, until another attempt is made to find a suitable host root.

But once the fungal hyphae sense the presence of a potential host, they branch and proliferate profusely; this probably increases the chance of contact with the root and subsequently infecting it. The branching phenomenon was first described more than 30 years ago⁵, and since then it has always seemed likely that the agent involved was a 'branching factor' released by roots. Indeed, root exudates have been shown to contain such a factor⁶. But its chemical structure has escaped identification, mostly because the concentration in root exudates is very low and the factor seemed to be rather unstable.

That changes with the report of Akiyama *et al.*¹, which describes the chemical structure of branching factor and constitutes a leap

forward in our understanding of plant–fungal symbioses. Akiyama and colleagues came up with a procedure for concentrating the factor, in which water containing root exudates of a plant called *Lotus japonicus* was pumped through a cartridge containing active charcoal. Branching factor was enriched on the charcoal surface, and ultimately enough of it could be isolated to allow determination of its chemical structure. Another essential element in the authors' success was the establishment of a sensitive bioassay that allowed the activity of branching factor towards fungi to be monitored during the purification procedure. The molecules identified are members of a group of sesquiterpene lactones, known as the strigolactones, that are active towards fungi at very low concentrations. Their relative instability seems to be a virtue, because a stable molecule would accumulate in the soil and so not convey positional information.

In another symbiotic relationship — that between legume roots and nitrogen-fixing, root-nodule bacteria — chemical communication also occurs before infection. But it seems that fungal and bacterial root symbionts detect different plant signalling molecules. Plant flavonoids activate the bacterial production of symbiotic signalling molecules known as nodulation factors⁷, but strigolactones are structurally and biosynthetically unrelated to flavonoids.

Arbuscular mycorrhizal fungi are exceptionally promiscuous with respect to the host plants they colonize; for example, my group has used a single strain of one species, *Glomus intraradices*, to successfully infect a wide range of plants, including liverworts, grasses (rice, wheat) and dicotyledonous plants such as tomato and pea. This is consistent with the broad distribution of strigolactones in the plant kingdom, and isolates from the two main groups of flowering plants, the dicotyledons and monocotyledons, do not significantly differ in their biological activity¹.

A twist in the story comes from the close relatedness between branching factor and molecules detected by parasitic weeds⁸. For example, the pest *Striga* (witchweed) causes devastating losses in crops, and is a major obstacle to food production in Africa. The roots of *Striga* and related plants can attack the roots of their victims and rob them of water and nutrients⁹. It seems that these parasites rely on the very same molecules for seed germination that facilitate symbiosis of roots with arbuscular mycorrhizal fungi. Interestingly, roots increase strigolactone production when phosphate availability is low, possibly to attract the fungi, and these are the very conditions under which *Striga* infestations occur¹⁰.

To understand molecular communication between plant and fungus, one of the most pressing questions is now the chemical structure of the fungal signals that induce symbiosis-related gene expression in plant roots.

Some observations suggest that the fungal signal is produced only by hyphae that have undergone the branching response — that is, after perception of branching factor¹¹. Akiyama and colleagues' results open up the possibility of inducing fungal responses in the absence of a host root, which might also help in identifying the fungal signal.

Despite the global importance of arbuscular mycorrhizal fungi and their potential in agriculture, research into their molecular genetics is still in its infancy. What deters geneticists who work with model systems, such as yeast, is that these fungi contain several nuclei within a single spore and the nuclei are probably genetically diverse¹². Together with a complete absence of sex in all arbuscular mycorrhizal fungi that have been investigated, this hinders classical genetic approaches. A programme to sequence the entire 15-megabase genome of *G. intraradices* is under

way — I hope the results will prompt more researchers to tackle the molecular biology of these fungi that are so essential to plant life on our planet.

Martin Parniske is at the Genetics Institute, Ludwig-Maximilians Universität, Maria-Ward-Strasse 1a, 80638 Munich, Germany.
e-mail: parniske@lmu.de

1. Akiyama, K., Matsuzaki, K.-i. & Hayashi, H. *Nature* **435**, 824–827 (2005).
2. Remy, W., Taylor, T. N., Hass, H. & Kerp, H. *Proc. Natl Acad. Sci. USA* **91**, 11841–11843 (1994).
3. Redecker, D., Kodner, R. & Graham, L. E. *Science* **289**, 1920–1921 (2000).
4. Brundrett, M. *New Phytol.* **154**, 275–304 (2002).
5. Mosse, B. *Annu. Rev. Phytopathol.* **11**, 171–196 (1973).
6. Buee, M. et al. *Mol. Plant Microbe Interact.* **13**, 693–698 (2000).
7. Lerouge, P. et al. *Nature* **344**, 781–784 (1990).
8. Cook, C. E. et al. *Science* **154**, 1189–1190 (1966).
9. Yoder, J. I. *Curr. Opin. Plant Biol.* **4**, 359–365 (2001).
10. Yoneyama, K., Takeuchi, Y. & Yokota, T. *Physiol. Planta* **112**, 25–30 (2001).
11. Kosuta, S. et al. *Plant Physiol.* **131**, 952–962 (2003).
12. Hijiri, M. & Sanders, I. R. *Nature* **433**, 160–163 (2005).

COMPUTATIONAL SCIENCE

Can get satisfaction

Carla P. Gomes and Bart Selman

The sheer complexity of some computational problems means they will probably never be solved, despite the ever-increasing resources available. But we can sometimes predict under what conditions solutions exist.

Computer scientists have been quite successful at developing fast algorithms: Google, for example, searches its index of more than eight billion web pages in a fraction of a second. The indexed-search problem is said to be 'tractable', or efficiently solvable; it is even possible to guarantee that, no matter what keywords you search on, you will get an answer quickly. Such a worst-case efficiency guarantee is the gold standard for algorithm design. But unfortunately, such a guarantee is not always possible: for many computational problems, we can

either prove that no efficient algorithm exists, or we have good evidence that such an algorithm is unlikely to exist. Achlioptas *et al.*¹ (page 759 of this issue) provide a series of rigorous insights into a class of computational puzzles — known as constraint-satisfaction problems — that occur in areas as diverse as the scheduling of airline and train crews, data-mining, software design and computational biology.

Constraint-satisfaction problems are defined on the one hand by a set of variables

(representing the unknowns of the problem), each of which can have a finite number of possible values, and on the other by a set of constraints. In constructing a schedule for a knockout sports tournament, for example, we can introduce a series of variables that can take one of two values: 1, 'true'; or 0, 'false'. The variable *X* might stand for 'team A plays team B in round 1'; if *X* is set to 1, this statement is true. If we introduce a second variable, *Y*, representing 'team A plays team C in round 1', we must then also introduce a constraint that says 'if *X* is 1, then *Y* should be 0'. This constraint is used to encode the fact that team A cannot play two different teams in the same round.

The encoding of real-world problems requires thousands of variables with tens of thousands of constraints. The challenge is to find an assignment of values to the variables such that all constraints are satisfied. With *N* binary (two-valued) variables, there are 2^N possible value assignments; so if *N* is 1,000, we have an enormous 'search-space' of $2^{1,000}$ — more than 10^{300} — different assignments, the vast majority of which violates one or more constraints. This exponentially scaling search-space is much too large to be examined explicitly. Computer scientists conjecture, however, that there is no algorithm that can do substantially better than an exhaustive search — and an algorithm that works well for all possible sets of constraints is thus unlikely to exist². But how difficult is it to find an assignment of values for a 'typical' set of constraints, such as might be found in a real-world application?

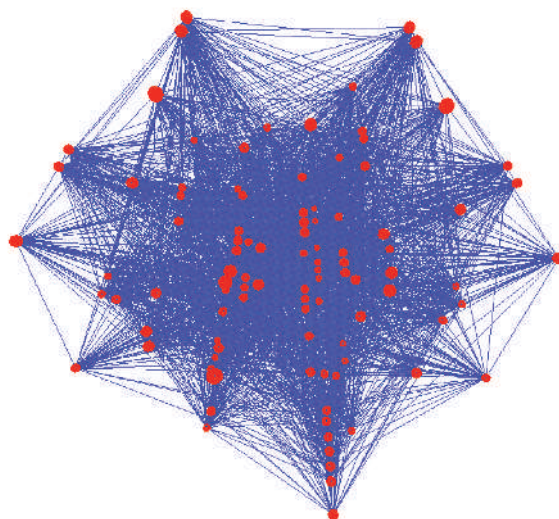
Certain sets of constraints are actually surprisingly easy to satisfy. In particular, if the problem has many variables with many possible values and only a few constraints, there will be many solutions that are relatively easy to find. On the other hand, if the problem contains few variables but a large number of constraints, often no assignment exists that satisfies all the constraints. The computationally interesting problems lie somewhere in the middle. For randomly generated constraint-satisfaction problems (Fig. 1), a sudden, sharp 'phase transition' occurs at a certain ratio of constraints to variables: at this value, we pass from a situation where almost all combinations of constraints can be satisfied, to one where very few combinations can be satisfied^{3–8}.

Achlioptas *et al.*¹ present a new approach to determining the lower bound on this phase transition in the so-called *k*-SAT problem. This is a constraint-satisfaction problem that has only binary variables and a special logical form of constraints that contain exactly *k* variables. It has been shown that any general constraint-satisfaction problem can be translated into a *k*-SAT problem, for any values of *k* larger than 2 (Fig. 2, overleaf). In fact, thousands of practical computational problems — from optical switching, supply-chain management and chip design, to protein folding — can be formulated as *k*-SAT problems.

Achlioptas and colleagues introduce a

Figure 1 | Web of constraint.

The dependency graph of a random constraint-satisfaction problem at a constraint-to-variable ratio of 7.9 (100 binary variables and 790 constraints). Each red dot represents a variable; a line connects two variables if they share a constraint. This graph arises from a problem in which each constraint connects exactly four variables (a 4-SAT problem). This problem is satisfiable — that is, there is an assignment to the variables such that all constraints are satisfied — as predicted by Achlioptas *et al.*¹. (Figure devised by Anand Kapur.)



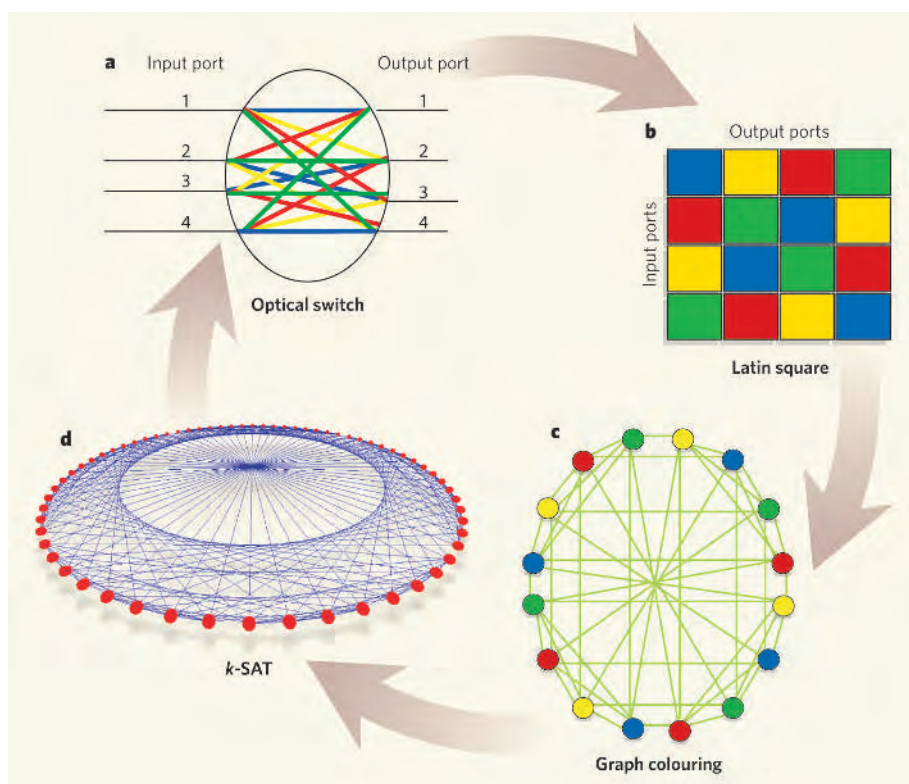


Figure 2 | Finding satisfaction. Constraint-satisfaction problems form a large class of practical computational problems, and encompass well-studied special cases such as the graph-colouring and k -SAT problems. **a**, An optical switch. The task is to assign a wavelength of light to each input–output path, such that a given wavelength is not assigned more than once to any input or output port. **b**, The optical-switch problem is equivalent to assigning different colours to the cells of a square matrix such that there is no repeated colour in any row or column (a ‘Latin square’). This problem can be translated into **c**, a graph-colouring problem or **d**, a k -SAT problem. Achlioptas *et al.*¹ introduce a technique for determining whether random k -SAT problems will have a satisfying assignment — that is, an assignment to the variables such that all constraints are satisfied — and whether a random graph can be coloured with a given number of colours.

sophisticated probabilistic argument, the weighted-second-moment method, to show that values for variables that satisfy all the constraint clauses can be assigned with high probability up to a constraint-to-variable ratio of $2^k \ln 2 - k$. Their method cleverly considers the overall statistical properties of the exponential search-space, without actually identifying any specific satisfying assignment. The lower limits that they find for the position of the phase transition at different values of k are very close to the best-known upper limits, especially for larger values of k , thus limiting significantly the range of possible values where the phase transition can occur.

The authors also consider the graph-colouring problem, another widely studied constraint-satisfaction task (Fig. 2). In this problem, nodes of a graph are coloured such that no two connected nodes have the same colour. (Ensuring that, on a map, no two countries of the same colour border each other is an example of this problem.) Using the second-moment method, Achlioptas *et al.* can predict how many colours are needed to fill in random graphs — without actually showing a valid colouring.

Achlioptas and colleagues’ approach was inspired in part by work that used advanced

techniques from statistical physics to obtain insights into the k -SAT phase transition⁹. In turn, physicists have used the second-moment method to show rigorously that, as the phase-transition region is approached, the exponential search-space fractures dramatically, with many small solution clusters appearing relatively far

apart from each other¹⁰. Traditional search procedures have trouble finding solutions in these search-spaces because they tend to get ‘stuck’ in between the solution clusters.

The quest for a deeper understanding of phase-transition phenomena in computational problems has been a catalyst to a productive interchange of ideas and concepts between statistical physics, mathematics and computer science. Step-by-step, this work is revealing the intricate structure of exponential search-spaces. These insights can lead to the design of new methods for searching such spaces⁹, and thus to fundamentally new algorithms for computational problems. An open question is whether such techniques can be adapted to deal with constraint-satisfaction problems that have more inherent structure¹¹ than the k -SAT problem, and which occur in many real-world applications.

Carla P. Gomes is in the Faculty of Computing and Information Science, Departments of Applied Economics and Management and of Computer Science, and Bart Selman is in the Department of Computer Science, Cornell University, 4148 Upson Hall, Ithaca, New York 14853, USA. e-mail: selman@cs.cornell.edu

1. Achlioptas, D., Naor, A. & Peres, Y. *Nature* **435**, 759–764 (2005).
2. Clay Mathematics Institute *The P versus NP problem* www.claymath.org/prizeproblems/pvsnp.htm
3. Cheeseman, P., Kanefsky, B. & Taylor, W. *Proc. 12th Int. Joint Conf. Artif. Intell.* 331–337 (Morgan Kaufmann, San Francisco, CA, 1991).
4. Mitchell, D., Selman, B. & Levesque, H. *Proc. 10th Natl Conf. Artif. Intell.* 459–465 (AAAI Press, Menlo Park, CA, 1992).
5. Kirkpatrick, S. & Selman, B. *Science* **264**, 1297–1301 (1994).
6. Hogg, T., Huberman, B. A. & Williams, C. (eds) *Frontiers in Problem Solving: Phase Transitions and Complexity spec. issue Artif. Intell.* **81** (1996).
7. Friedgut, E. *J. Am. Math. Soc.* **12**, 1017–1054 (1999).
8. Monasson, R., Zecchina, R., Kirkpatrick, S., Selman, B. & Troyansky, L. *Nature* **400**, 133–137 (1999).
9. Mezard, M., Parisi, G. & Zecchina, R. *Science* **297**, 812–815 (2002).
10. Mezard, M., Mora, T. & Zecchina, R. preprint at <http://arxiv.org/cond-mat/0504070> (2005).
11. Williams, R., Gomes, C. & Selman, B. *Proc. 17th Int. Joint Conf. Artif. Intell.* 1173–1178 (Morgan Kaufmann, San Francisco, CA, 2003).

CANCER

Inflammation by remote control

Alberto Mantovani

Smouldering beneath many latent tumours is a chronic inflammation that goads pre-malignant cells into becoming full-blown cancer. The spark that kindles these flames comes from an unexpected source.

Chronic inflammation makes individuals susceptible to many forms of cancer. The culprits that drive this process are inflammatory cells and signalling molecules of the ‘innate’ immune system — our in-born defence system, which recognizes potential threats without previous exposure to them. But in a surprising twist, de Visser *et al.*, writing in

*Cancer Cell*¹, demonstrate that specialized cells from the ‘adaptive’ immune system orchestrate the innate inflammation that promotes tumour progression.

The link between inflammation and the promotion of cancer was first observed in the nineteenth century², but only in recent years has it become a generally accepted

phenomenon²⁻⁴. Epidemiological studies have shown that chronic inflammation predisposes individuals to certain cancers, and conversely that non-steroidal anti-inflammatory agents protect against several tumours. Most pre-cancerous and cancerous tissues show signs of inflammation; this involves the movement of innate immune cells into the tissue, the presence of specific inflammatory signalling molecules (called cytokines and chemokines), changes in tissue structure (remodelling), and the formation of new blood vessels (angiogenesis).

Further studies found that cancer-associated inflammation actually promotes tumour growth and progression²⁻⁶. For instance, innate immune cells called tumour-associated macrophages work their way into precancerous tissue, and can release factors that promote tumour growth and metastasis^{2,7}. Accordingly, in many human tumours, the infiltration of large numbers of these macrophages is associated with poor prognosis. Moreover, increased expression of genes associated with macrophage infiltration (such as CD68) forms part of the molecular signatures that herald poor prognosis in certain cancers⁸.

The best-studied examples of inflammation-associated cancer are colon cancer and cervical carcinoma. Cervical carcinoma is caused by infection with human papilloma virus (HPV). Genetically engineered mice that carry some of the genes of HPV strain 16 — targeted to be expressed in the skin epithelial

tissue that gives rise to carcinomas — recapitulate many of the stages of this disease. Innate immune cells, most prominently mast cells and granulocytes, infiltrate the pre-malignant epithelial tissues, and they are followed by macrophages as the cancer develops further (Fig. 1). Innate immune cells drive a chronic inflammatory process that promotes overproliferation of epithelial cells, tissue remodelling and angiogenesis, followed eventually by invasive carcinoma⁹.

Using the HPV16 mice, de Visser *et al.*¹ examined the interplay between innate and adaptive immunity in the progression of these tumours. They bred the HPV16 mice with mice that lack both types of adaptive immune cell — T cells and B cells. This genetic elimination of adaptive cells blocked the recruitment of innate cells, and the subsequent tissue remodelling and angiogenesis. Carcinogenesis was therefore arrested at the stage of epithelial overproliferation. When B cells from HPV16 mice were transferred into the immune-deficient HPV16 cross-breed animals, the expected inflammatory response and cancer progression were restored, showing that these cells are the perpetrators.

B cells do not infiltrate the precancerous tissues, and so they must orchestrate the innate immune cells remotely. This was confirmed by the finding that blood serum (without cells) from HPV16 mice also reinstated cancer progression in the HPV16 cross-breed mice. But what is the remote-control mechanism that

links B cells with the innate immune response? One possibility is antibodies, which are produced by B cells. Indeed, de Visser *et al.* found preliminary circumstantial evidence that antibodies may drive the cancer-promoting inflammation (Fig. 1).

These results are surprising and provocative. In a context in which adaptive immunity promotes inflammation, one would expect T cells to be the prime mover, but this is clearly not the case here. Moreover, de Visser *et al.* found that the B-cell antibodies, rather than occurring next to the overproliferating epithelial cells, were in fact found in the underlying dermis layer. This argues against the possibility that the antibodies trigger immune responses by recognizing molecules on the surface of the tumour cells. The extracellular milieu surrounding tumours can contain peculiar components^{10,11}, and de Visser *et al.* speculate that cancer-associated changes in this milieu may elicit the antibody response that kindles the innate immune response. Immune complexes in the local tumour environment would then activate innate immune cells to sustain chronic inflammation and further tumour progression.

Overall, de Visser and colleagues draw attention to the poorly explored issue of the interplay between the innate and the adaptive arms of immunity — on the one hand in surveillance against cancer by T cells¹², and on the other in inflammation-driven cancer promotion. It will be necessary to show definitively that antibodies are indeed the molecular messengers that allow B cells to control the innate immune response, and to define their specificity. Moreover, whether B cells have a general role in controlling the chronic inflammation associated with other types of cancer, and the mechanism involved (for example, requirement for T-cell help and molecules recognized by antibodies), should also be assessed. Therapeutic targeting of cancer-promoting inflammatory reactions is in the early stages of development⁴, and its progress will depend on defining the underlying cellular and molecular mechanisms in the relevant systems. ■

Alberto Mantovani is at the Istituto Mario Negri, Via Eritrea 62, 20157 Milan, and the ICH University of Milan, Milan, Italy.
e-mail: mantovani@marionegri.it

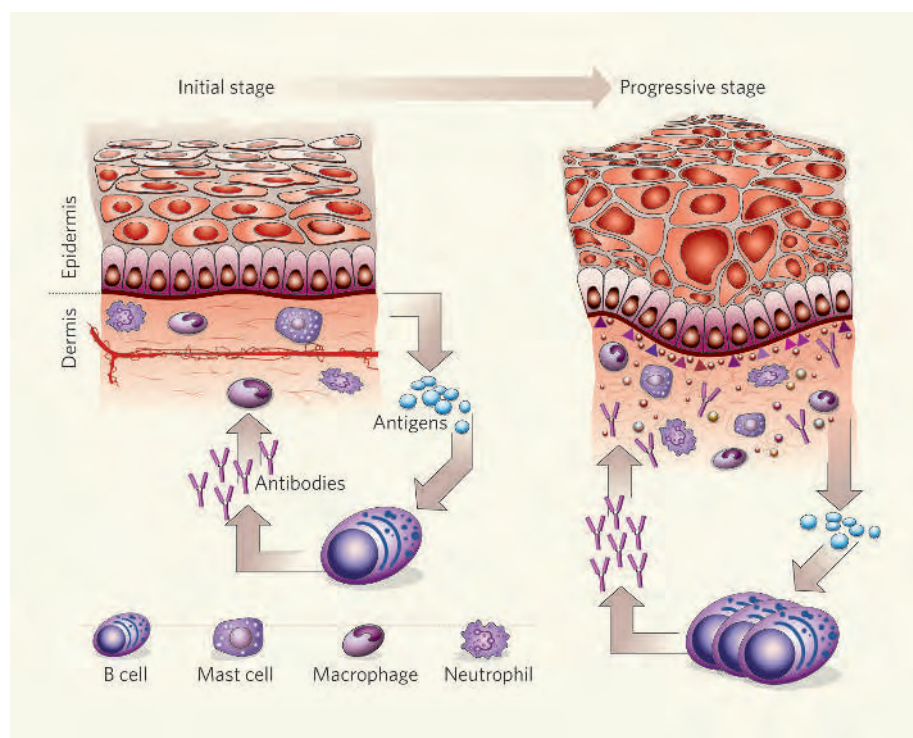


Figure 1 | Setting up inflammation in skin cancer. Work by de Visser *et al.*¹ shows that adaptive immune cells called B cells produce the signal that kindles inflammation in the epidermal layer of the skin. One model of how this might work is that a molecule (antigen) present in the tumour environment stimulates the B cells to release antibodies that target innate immune cells (such as neutrophils, macrophages and mast cells), which set up and maintain inflammation. Once activated, the innate cells release factors (triangles) that can promote growth and spreading of the tumour.

- de Visser, K., Korets, L. V. & Coussens, L. M. *Cancer Cell* **7**, 411–423 (2005).
- Balkwill, F. & Mantovani, A. *Lancet* **357**, 539–545 (2001).
- Coussens, L. M. & Werb, Z. *Nature* **420**, 860–867 (2002).
- Balkwill, F., Charles, K. A. & Mantovani, A. *Cancer Cell* **7**, 211–217 (2005).
- Gretchen, F. R. *et al. Cell* **118**, 285–296 (2004).
- Pikarsky, E. *et al. Nature* **431**, 461–466 (2004).
- Mantovani, A., Sozzani, S., Locati, M., Allavena, P. & Sica, A. *Trends Immunol.* **23**, 549–555 (2002).
- Paik, S. *et al. N. Engl. J. Med.* **351**, 2817–2826 (2004).
- Coussens, L. M., Tinkle, C. L., Hanahan, D. & Werb, Z. *Cell* **103**, 481–490 (2000).
- Borsi, L. *et al. Blood* **102**, 4384–4392 (2003).
- Sangaletti, S., Stoppacciaro, A., Guiducci, C., Torrisi, M. R. & Colombo, M. P. *J. Exp. Med.* **198**, 1475–1485 (2003).
- Dunn, G. P., Old, L. J. & Schreiber, R. D. *Immunity* **21**, 137–148 (2004).

BRIEF COMMUNICATIONS

Red tides and marine mammal mortalities

Unexpected brevetoxin vectors may account for deaths long after or remote from an algal bloom.

Potent marine neurotoxins known as brevetoxins are produced by the 'red tide' dinoflagellate *Karenia brevis*. They kill large numbers of fish and cause illness in humans who ingest toxic filter-feeding shellfish or inhale toxic aerosols¹. The toxins are also suspected of having been involved in events in which many manatees and dolphins died, but this has usually not been verified owing to limited confirmation of toxin exposure, unexplained intoxication mechanisms and complicating pathologies^{2–4}. Here we show that fish and seagrass can accumulate high concentrations of brevetoxins and that these have acted as toxin vectors during recent deaths of dolphins and manatees, respectively. Our results challenge claims that the deleterious effects of a brevetoxin on fish (ichthyotoxicity) preclude its accumulation in live fish, and they reveal a new vector mechanism for brevetoxin spread through food webs that poses a threat to upper trophic levels.

In the spring of 2002, 34 endangered Florida manatees (*Trichechus manatus latirostris*) died in southwest Florida, and 107 bottlenose dolphins (*Tursiops truncatus*) died in waters off the Florida panhandle in the spring of 2004. In both of these unusual mortality events, extensive water surveys revealed that only low concentrations of *K. brevis* were present.

We tested for the presence of brevetoxin in the fluids and tissues of 63 of these animals (27 manatees, 36 dolphins) and found very high concentrations in the tissues of all of them (see supplementary information), confirming that the animals must have been exposed to brevetoxin. In a previous event, in which 149 manatees died, lung pathology indicated that brevetoxins had been inhaled³. In our examples, the absence of similar pathology excluded the possibility of poisoning through aerosol exposure, and the high toxin concentrations measured in the stomach contents indicated that the toxin was from a dietary source.

Manatee stomach contents were composed exclusively of seagrass; filter-feeding tunicates, which were suspected vectors in a 1982 mortality event³, were notably absent. Analysis of seagrass (*Thalassia testudinum*) collected at several locations in the area of death revealed high concentrations of brevetoxins (Fig. 1a), mainly in the epiphytic fraction (epiphytes, 83% of total brevetoxins; blades, 7%; rhizomes, 10%). The accumulation mechanism could involve active uptake or passive adsorption of

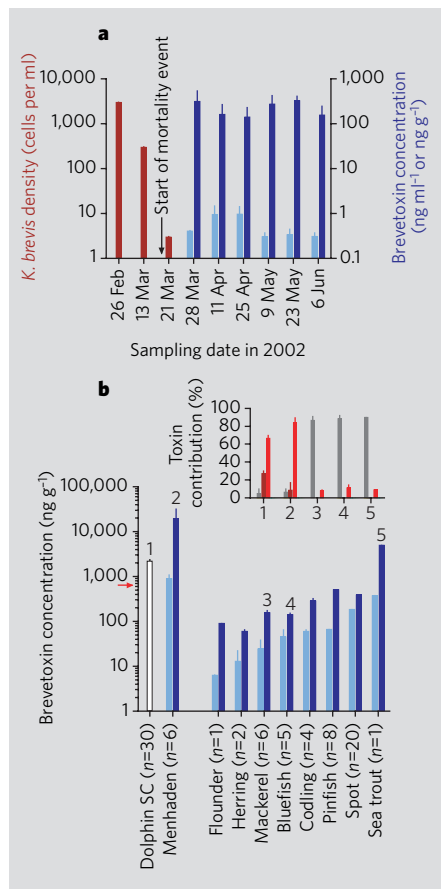


Figure 1 | Brevetoxin concentrations in seagrass and fish during mass-mortality events.

a, Density of red-tide algae *Karenia brevis* (red bars) and brevetoxin concentrations in seagrass (*Thalassia testudinum*) (dark blue bars; ng g⁻¹) and in sea water (light blue bars; ng ml⁻¹), collected during and after the 2002 manatee mortality event in Charlotte Harbor, Florida. Error bars, standard deviation between samples collected from four sites. **b**, Brevetoxin concentrations in dolphin stomach contents (SC), in undigested menhaden, and in fish collected live (light blue, in muscle; dark blue, in viscera) (flounder, *Paralichthys lethostigma*; herring, *Opisthonema oglinum*; mackerel, *Scomberomorus maculatus*; bluefish, *Pomatomus saltatrix*; codling, *Urophycis floridana*; pinfish, *Lagodon rhomboides*; spot, *Leiostomus xanthurus*; and sea trout, *Cynoscion nebulosus*) from St Joseph Bay, Florida, in spring 2004. Red arrow, regulation limit for brevetoxin in shellfish. Error bars, standard deviation between individual fish, except for pinfish and spot (pooled). Inset, toxins identified by liquid chromatography and mass spectroscopy in selected samples, as numbered in the main bar chart; those in dolphin stomach contents and in menhaden (1, 2) differed from the profile found in fish collected live two weeks after the onset of the mortality (3–5). Bars: dark red, brevetoxin-2; light red, brevetoxin-3; grey, brevetoxin-2 disulphide metabolite. For methods, see supplementary information.

the toxin. As the red tide that previously affected the area had almost dissipated by the start of the mortality event (Fig. 1a), the comparable toxin concentrations in manatee stomach contents and in the seagrass beds (up to 1,136 and 1,263 ng brevetoxin per g, respectively) indicated that seagrass was the primary source of brevetoxin for the manatees.

An extensive pathological, pathogenic and environmental investigation conducted in response to the dolphin mortalities failed to identify any consistent mortality factor other than brevetoxin⁶. Although no *K. brevis* was evident at the time, the contents of the dolphins' stomachs were acutely toxic. Stomachs were full and menhaden (*Brevoortia* spp.), a type of plankton-eating fish, were identified as the dominant prey in 50% of the 28 animals examined. Surprisingly, there was a high level of brevetoxin contamination in all undigested menhaden tested and, to a lesser extent, in all fish that were collected live two weeks after the

onset of the dolphin deaths (Fig. 1b).

Until now, it was uncertain whether live fish could accumulate and transfer brevetoxins to upper trophic levels, as brevetoxins kill fish even at low concentrations⁷ and typically result in high fish mortalities during red tides¹. To determine how brevetoxins might accumulate in fish, we exposed omnivorous and planktivorous fish to toxic shellfish (which retain brevetoxins after blooms have dissipated¹) and to bloom concentrations of healthy *K. brevis* cultures with low extracellular toxin concentrations (as sometimes observed during red tides⁸), respectively. We found that brevetoxin accumulates in both types of feeder (results not shown). Because brevetoxins are sequestered in their food (shellfish and *K. brevis* cells), the fish remained healthy while brevetoxin concentrations increased in their tissues (up to 2,675 ng g⁻¹ in the viscera and 1,540 ng g⁻¹ in the muscle of omnivorous fish exposed for two weeks to toxin-containing clams).



Florida manatees (3 metres long, on average) are susceptible to toxins from the red tide alga *Karenia brevis* (inset; cell diameter, 30–35 mm).

Brevetoxin poisoning in humans has so far been restricted to the consumption of contaminated shellfish (neurotoxic shellfish poisoning). Although the accumulation of brevetoxins in live fish to the levels measured in the menhaden is probably short-lived and unusual, this finding, together with the dolphin deaths (given that dolphins are a sentinel species⁹), raises concerns that humans could also be poisoned by contaminated fish.

These findings show not only that brevetoxin-contaminated food webs pose a threat to marine mammals, but also that toxin vectors can result in delayed or remote animal exposure. Biological toxins should therefore be considered as possible culprits when investigating unusual marine animal mortalities, even in the absence of toxin-producing algae.

Leanne J. Flewelling*, **Jerome P. Naar†**,
Jay P. Abbott*, **Daniel G. Baden†**, **Nélio B. Barros‡**,
Gregory D. Bossart§, **Marie-Yasmine D. Bottein¶**,
Daniel G. Hammond*, **Elsa M. Haubold***,
Cynthia A. Heil*, **Michael S. Henry‡**,
Henry M. Jacocks†, **Tod A. Leighfield¶**,
Richard H. Pierce‡, **Thomas D. Pitchford***,
Sentiel A. Rommel*, **Paula S. Scott***,
Karen A. Steidinger*, **Earnest W. Truby***,
Frances M. Van Dolah¶, **Jan H. Landsberg***

*Florida Fish and Wildlife Conservation Commission, Fish and Wildlife Research Institute, St Petersburg, Florida 33701, USA

†Center for Marine Science, University of North Carolina, Wilmington, North Carolina 28409, USA
e-mail: naarj@uncw.edu

‡Mote Marine Laboratory, Sarasota, Florida 34236, USA

§Harbor Branch Oceanographic Institution, Fort Pierce, Florida 34946, USA

¶NOAA, NOS, Center for Coastal Environmental Health and Biomolecular Research, Charleston, South Carolina 29412, USA

1. Steidinger, K. A. in *Algal Toxins in Seafood and Drinking Water* (ed. Falconer, I.) 1–28 (Academic, London, 1993).
2. Geraci, J. R. *Clinical Investigation of the 1987–1988 Mass Mortality of Bottlenose Dolphins along the US Central and South Atlantic Coast*. Report to the National Marine Fisheries Service and US Navy, Office of Naval Research and Marine Mammal Commission (Ontario Veterinary College, Guelph, Ontario, 1989).
3. O'Shea, T. J., Bonde, R. K., Buerge, C. D. & Odell, D. K. *Mar. Mamm. Sci.* **7**, 165–179 (1991).
4. Van Dolah, F. M., Doucette, G. J., Gulland, F. M. D., Rowles, T. L. & Bossart, G. D. in *Toxicology of Marine Mammals* (eds Vos, J. G., Bossart, G. D., Fournier, M. & O'Shea, T. J.) 247–269 (Taylor & Francis, New York, 2003).
5. Bossart, G. D., Baden, D. G., Ewing, R. Y., Roberts, B. & Wright, S. D. *Toxicol. Pathol.* **26**, 276–282 (1998).
6. Working Group on Marine Mammal Unusual Mortality Events *The Bottlenose Dolphin (Tursiops truncatus): Unusual Mortality Event along the Panhandle of Florida March–April 2004* (National Marine Fisheries Service, 2004).
7. Baden, D. G. & Mende, T. J. *Toxicon* **20**, 457–461 (1982).
8. Pierce, R. H., Henry, M. S., Blum, P. & Payne, S. in *Harmful Algal Blooms 2000* (eds Hallegraeff, G. M. et al.) 421–424 (Intergovernmental Oceanic Commission, Paris, 2001).
9. Ross, P. S. *Hum. Ecol. Risk Ass.* **6**, 29–46 (2000).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.
doi:10.1038/nature435755a

SEISMOLOGY

Earthquake risk on the Sunda trench

On 28 March 2005 the Sunda megathrust in Indonesia ruptured again, producing another great earthquake three months after the previous one. The rupture was contiguous with that of the December 2004 Sumatra–Andaman earthquake, and is likely to have been sparked by local stress, although the triggering stresses at its hypocentre were very small — of the order of just 0.1 bar. Calculations show that stresses imposed by the second rupture have brought closer to failure the megathrust immediately to the south, under the Batu and Mentawai islands, and have expanded the area of increased stress on the Sumatra fault. Palaeoseismologic studies show that the Mentawai segment of the Sunda megathrust is well advanced in its seismic cycle and is therefore a good candidate for triggered failure.

The 1,300-km-long Sumatra–Andaman rupture of the Sunda megathrust that occurred on 26 December 2004 shed stresses on to other structures in the region. We previously identified two faults of particular concern: the continuation of the Sunda megathrust to the south, beneath the islands of Simeulue and

Table 1 | Hypocentral stresses in the Sumatran earthquake of 28 March 2005

	Slip distribution	
	Ref. 8	Ref. 9
Co-seismic stress	0.005	0.110
Post-seismic stress	0.064	0.060
Total stress	0.069	0.170

Hypocentral stresses are shown in bars and are calculated using wave-form slip inversions^{8,9} and an oceanic earth structure after ref. 10. The difference in co-seismic stresses is largely due to the difference in the southward extent of the 26 December rupture in the two slip models.

Nias, and the vertical, strike-slip Sumatra fault¹. On 28 March, rupture of the Simeulue–Nias segment generated a magnitude-8.7 earthquake, which caused widespread destruction on the islands and is estimated to have killed about 2,000 people.

We have calculated the stresses induced by the Sumatra–Andaman rupture at the hypocentre of the Simeulue–Nias earthquake, including both the co-seismic² elastic effect and the effect of post-seismic³ viscoelastic relaxation of the upper mantle. The total stress perturbation at the hypocentre was small

(Table 1), between 0.07 and 0.17 bars. The size of this triggering stress illustrates the extreme non-linearity of the earthquake nucleation process.

Like its predecessor, the Simeulue–Nias earthquake has appreciably altered the state of stress in the surrounding region (Fig. 1a). Although it changed only slightly the level of stress on the section near Banda Aceh, which was most affected by the Sumatra–Andaman rupture, it has increased stresses on the Sumatra fault south of that section. As in the case of the Sumatra–Andaman rupture, the section of the megathrust just to the south has also been stressed appreciably (by as much as 8 bars on the section beneath the Batu islands and somewhat less on the segment beneath the Mentawai islands). Despite the smaller size of the Simeulue–Nias event, the magnitude of its stress perturbation on the Batu and northern Mentawai sections of the megathrust is similar to that which triggered the Simeulue–Nias earthquake. This stress may be expected to migrate further south over time as a result of viscoelastic effects.

The Batu section of the fault (Fig. 1b), from the Equator to about 0.7° S, last ruptured in 1935 during a magnitude-7.7 earthquake that resulted in about 2.3 metres of slip on a 70 km × 35 km patch of the megathrust^{4,5}. Recent palaeogeodetic studies (manuscript

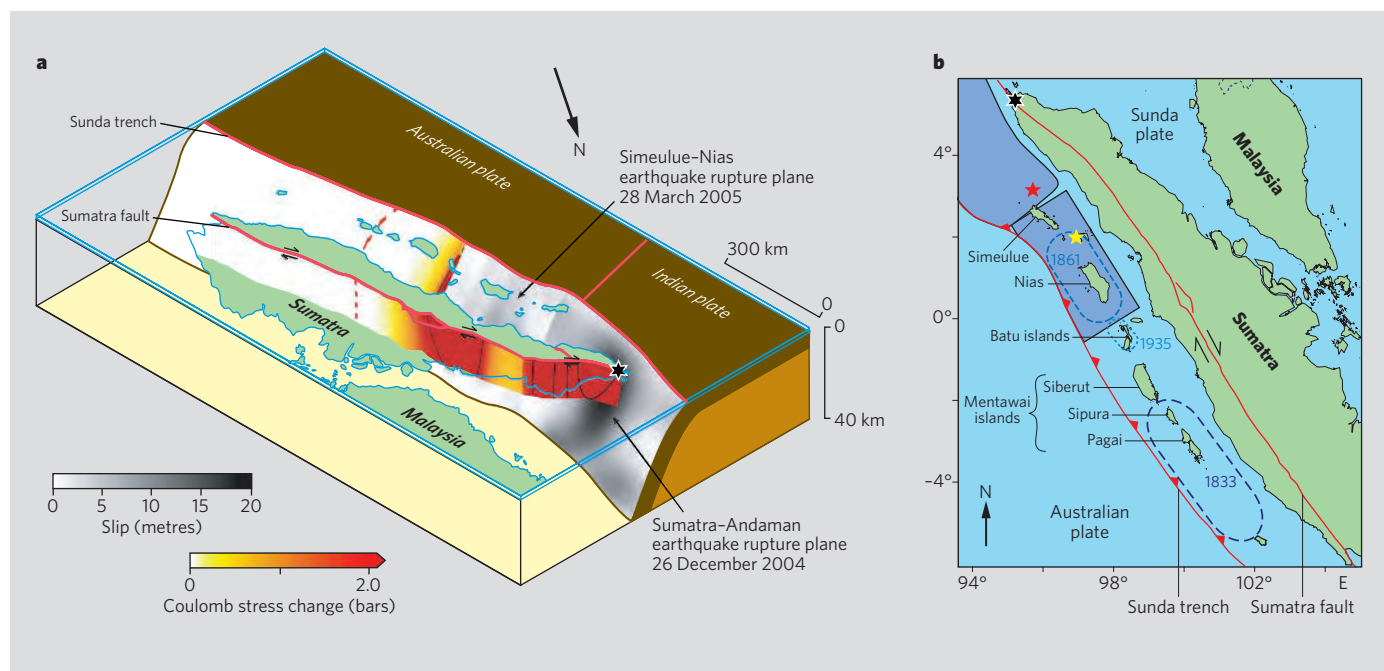


Figure 1 | The faults around Sumatra and the Sunda trench. **a**, Schematic of the Sumatran subduction zones with the overlying plates removed. Calculated three-dimensional stresses, including contributions from both earthquakes resolved directly on to the structures of interest, have been projected on to a diagram of the structural geometry and geography of the region. Here we use a low-value coefficient of effective friction, 0.4, although the main results are robust to large ranges of this value. Grey-scale values on the rupture plane represent the amount of slip in metres experienced on the southernmost 450 km of the Sumatra–Andaman earthquake and on the Simeulue–Nias earthquake. Colour-scale values represent the co-seismic stress change on the Sunda-trench subduction zone and the Sumatra fault. Stress contours are in 2-bar intervals. Red dashed contours indicate zero co-seismic stress. Black star indicates the location of Banda Aceh. **b**, Locations of ruptures of recent and historical earthquakes on the Sunda trench. Dotted lines indicate approximate extents of historical ruptures (1833, 1861 and 1935); solid lines surrounding dark-blue areas indicate seismological inversions of recent earthquakes; red star, epicentre of December 2004 event; yellow star, epicentre of March 2005 event; black star, Banda Aceh. The rupture area of the 1797 event, which is not shown here, probably overlaps significantly with the 1833 event under Sipura and Pagai Islands and may extend under Siberut Island. The precise extent of this event strongly influences the estimated slip deficit on the megathrust.

submitted) show that the megathrust is slipping aseismically both above and below this narrow patch. The slippage occurs at the rate of plate convergence. Furthermore, the 1935 patch (Fig. 1b) has been slipping during the past century at about half the rate at which the plate is moving. Therefore, accumulated strains and hence stresses on the Batu patch are probably low. The Mentawai segment, on the other hand, presents a greater threat.

Our palaeoseismic investigations (manuscript submitted) show that the megathrust has not ruptured under the island of Siberut (0.7 to 2° S) since 1797. The latest ruptures farther south involved a few metres of slip in 1797 (magnitude > 8; 2.0–3.5° S) and a 10-metre rupture in 1833 (magnitude > 8.5; 2.0–5.5° S)⁶. Both of these events produced large tsunamis on the islands and mainland coast⁷. Events similar to 1833 seem to occur every 230 years on average.

These observations, in conjunction with the stress changes mentioned, suggest that the greatest current seismic threat from the Sunda megathrust comes from the Mentawai section, between about 0.7 and 5.5° S. Slip on the northern part of this section could be greater than 10 metres, depending on the timing of the last rupture (based on a convergence rate of 5 cm yr⁻¹). Slip on the southern portion could be as great as in 1833: that is, up to

10 metres. The historical record and the experience of the Sumatra–Andaman and Simeulue–Nias events indicate that a tsunami could be a possibility. The threat of an earthquake of magnitude 7.0–7.5 on the Sumatra fault north of 4° N has not receded.

Suleyman S. Nalbant*, **Sandy Steacy***, **Kerry Sieh†**, **Danny Natawidjaja†**, **John McCloskey***

*School of Environmental Sciences, University of Ulster, Coleraine, County Derry BT52 1SA, UK
e-mail: j.mccloskey@ulster.ac.uk

†Tectonic Observatory, California Institute of Technology, Pasadena, California 91125, USA

1. McCloskey, J., Nalbant, S. S., & Steacy, S. *Nature* **434**, 291 (2005).

- Stein, R. S., Barka, A. A. & Dieterich, J. H. *Geophys. J. Int.* **128**, 594–604 (1997).
- Pollitz, F. F. & Sacks, I. S. *Bull. Seismol. Soc. Am.* **87**, 1–10 (1997).
- Rivera, L. et al. *Bull. Seismol. Soc. Am.* **92**, 1721–1736 (2002).
- Natawidjaja, D. H. et al. *J. Geophys. Res.* **109**, B04306 (2004).
- Siehl, K., Stebbins, C., Natawidjaja, D. H. & Suwargadi, B. W. *EOS Trans. AGU* **85**, Abstr. PA23A-1444 (2005).
- Newcomb, K. R. & McCann, W. R. *J. Geophys. Res.* **92**, 421–439 (1987).
- Ji, C. <http://www.gps.caltech.edu/~jichen/Earthquake/2005/sumatra/sumatra.html> (2005).
- Yagi, Y. <http://iisee.kenken.go.jp/staff/yagi/eq/Sumatra2004/Sumatra2004.html> (2005).
- Pollitz, F. F., Burgmann, R. & Romanowicz, B. *Science* **280**, 1245–1249 (1998).

Competing financial interests: declared none.
doi:10.1038/nature435756a

PLANT BIOCHEMISTRY

Anthocyanin biosynthesis in roses

Anthocyanin is the principal pigment in flowers, conferring intense red-to-blue cyanic colours on petals and helping to attract pollinators. Its biosynthesis involves glycosylation steps that are important for the stability of the pigment and for its aqueous solubility in vacuoles^{1,2}. Here we describe anthocyanin biosynthesis in roses (*Rosa hybrida*), which is unlike

the pathway used in other flowers in that it relies on a single enzyme to achieve glycosylation at two different positions on the precursor molecule. Phylogenetic analysis also indicates that this previously unknown glucosyltransferase enzyme may be unique to roses, with glycosylation having apparently evolved into a single stabilizing step in other plants.

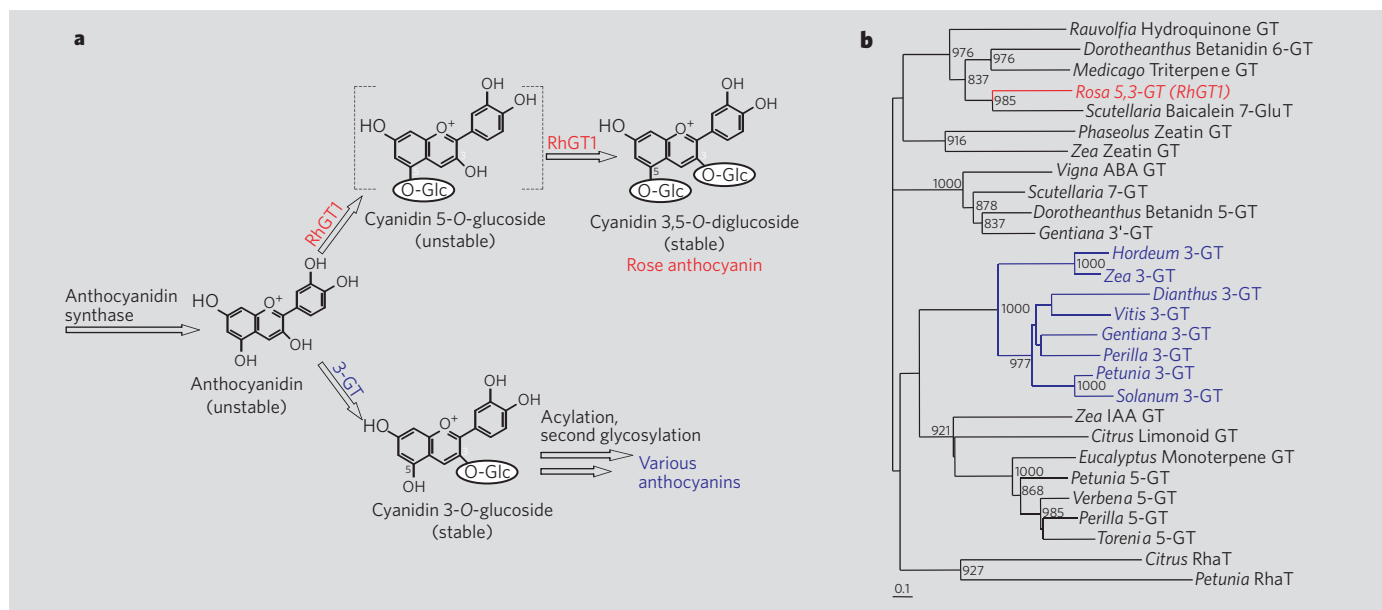


Figure 1 | A previously undiscovered rose anthocyanidin glucosyltransferase and its phylogeny. **a**, Comparison of the reaction pathways of anthocyanin glycosylation in the rose and in other plants: the rose glucosyltransferase RhGT1 catalyses two reactions instead of one. 3-GT, glucosyltransferase specific for the hydroxyl group at position 3 on the anthocyanidin molecule. **b**, Molecular phylogenetic tree based on the amino-acid sequences of members of the plant glucosyltransferase superfamily. Sequences were aligned by using the multiple alignment program Clustal W; the tree was constructed by the neighbour-joining method. Numbers indicate bootstrap values greater than 800.

More than 24,000 cultivars of roses have been registered^{3,4}. Although the flowers come in a wide variety of colours, rose pigments are composed mainly of the structurally simple anthocyanidin 3,5-O-diglucoside (anthocyanin)⁵. However, the process by which anthocyanidin is glycosylated in roses is poorly understood. We therefore investigated the role of glucosyltransferase activity in anthocyanin biosynthesis in rose petals (for methods, see supplementary information).

We isolated glucosyltransferase activity from rose petals and found that it could catalyse glycosylation at not just one but two different sites on the anthocyanidin molecule: glycosylation occurs first at the 5-OH and then at the 3-OH group (Fig. 1a). No new products were detected in reactions in which anthocyanidin 3-O-glucosides were used as glucosyl acceptors. These findings contradict the assumption^{2,6–8} that the stabilization of anthocyanidin occurs by initial glycosylation of the 3-OH residue of anthocyanidin.

This dual glucosyltransferase activity of the crude enzyme fraction was compared with that of a recombinant enzyme produced by cloning its encoding complementary DNA (RhGT1; accession number AB201048; length 1,419 base pairs). The recombinant RhGT1 protein also sequentially catalyses two glycosylation steps at the 5-OH and 3-OH positions (Fig. 1a, top). Like the crude enzyme, RhGT1 can use either unglycosylated anthocyanidin or anthocyanidin 5-O-glucoside as a glucosyl acceptor, but not anthocyanidin 3-O-glucoside.

Recombinant 3-glucosyltransferases that glycosylate flavonoids such as anthocyanidin in other plants (as, for example, in the carnation *Dianthus caryophyllus*) do not recognize antho-

cyanidin 5-O-glucoside as a glucosyl acceptor⁹, and recombinant 5-glucosyltransferases do not glycosylate anthocyanidin^{8,10}. These enzymes evidently differ from the rose enzyme RhGT1, which is a single enzyme that can catalyse glycosylation at two different positions on anthocyanidin. We designate this novel glucosyltransferase as UDP-glucose:anthocyanidin 5,3-O-glucosyltransferase, in which uridine diphosphate acts as the glucose carrier.

We used northern blotting to test for the presence of RhGT1 RNA in different parts of the rose (see supplementary information) and found that it is temporally and spatially regulated during flower development (results not shown), depending on when and where anthocyanin biosynthesis is activated for cyan coloration of rose flowers.

Molecular phylogenetic analysis (Fig. 1b) reveals that RhGT1 differs significantly from other anthocyanidin 3- and 5-glucosyltransferase subfamily members and that it probably evolved more closely with the glycosylation enzymes of pigments other than anthocyanin or of their secondary metabolites. It is a mystery why this particular glucosyltransferase evolved independently in roses. The novelty of the RhGT1 enzyme therefore lies not only in its ability to catalyse glycosylation at two different sites on the anthocyanidin molecule but also in its apparent absence from other species.

Like cyanidin, anthocyanidin 5-O-glucoside is unstable without the additional glycosylation at its 3-OH residue and so does not exist as a stoichiometric intermediate. In the rose pathway, anthocyanidin 3,5-O-diglucoside is therefore the first stable anthocyanin, whereas this is usually anthocyanidin 3-O-glucoside in other plants (Fig. 1a)^{2,6–8}. Antho-

cyanidin 3,5-O-diglucoside and anthocyanidin 3-O-glucoside are therefore responsible for flower coloration in roses and in other plants, respectively. Although many angiosperms produce anthocyanin derivatives from anthocyanidin 3-O-glucoside as a precursor^{2,8}, this is evolutionarily precluded in roses by their different glycosylation pattern, which may be unique to members of the Rosaceae.

Jun Ogata*, Yoshiaki Kanno*, Yoshio Itoh†, Hidehito Tsugawa*, Masahiko Suzuki*

*Aomori Green BioCenter, Nogi-Yamaguchi, Aomori 030-0142, Japan

e-mail: masahiko_suzuki@ags.pref.aomori.jp

†Department of Biotechnology, Faculty of Technology, Tokyo University of Agriculture and Technology, Koganei, Tokyo, 184-8588, Japan

1. Strack, D. & Wray, V. in *The Flavonoids. Advances in Research Since 1986* (ed. Harborne, J. B.) 1–22 (Chapman & Hall, London, 1993).
2. Heller, W. & Forkmann, G. in *The Flavonoids. Advances in Research Since 1986* (ed. Harborne, J. B.) 499–535 (Chapman & Hall, London, 1993).
3. Gudim, S. in *Encyclopedia of Rose Science* (eds Roberts, A. V., Debener, T. & Gudim, S.) 1, 25–30 (Elsevier, Amsterdam, 2003).
4. Cairns, T. in *Modern Roses XI: The World Encyclopedia of Roses* (Academic, San Diego, 2000).
5. Jay, M. et al. in *Encyclopedia of Rose Science* (eds Roberts, A. V., Debener, T. & Gudim, S.) 1, 248–255 (Elsevier, Amsterdam, 2003).
6. Tanaka, Y., Katsumoto, Y., Braglieri, F. & Mason, J. *Plant Cell Tiss. Organ Culture* **80**, 1–24 (2005).
7. Holton, T. A. & Cornish, E. C. *Plant Cell* **7**, 1071–1083 (1995).
8. Springob, K., Nakajima, J., Yamazaki, M. & Saito, K. *Nat. Prod. Rep.* **20**, 288–303 (2003).
9. Ogata, J., Itoh, Y., Ishida, M., Yoshida, H. & Ozeki, Y. *Plant Biotechnol.* **21**, 367–375 (2004).
10. Yamazaki, M. et al. *J. Biol. Chem.* **274**, 7405–7411 (1999).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.
doi:10.1038/nature43575a

Rigorous location of phase transitions in hard optimization problems

Dimitris Achlioptas¹, Assaf Naor¹ & Yuval Peres²

It is widely believed that for many optimization problems, no algorithm is substantially more efficient than exhaustive search. This means that finding optimal solutions for many practical problems is completely beyond any current or projected computational capacity. To understand the origin of this extreme 'hardness', computer scientists, mathematicians and physicists have been investigating for two decades a connection between computational complexity and phase transitions in random instances of constraint satisfaction problems. Here we present a mathematically rigorous method for locating such phase transitions. Our method works by analysing the distribution of distances between pairs of solutions as constraints are added. By identifying critical behaviour in the evolution of this distribution, we can pinpoint the threshold location for a number of problems, including the two most-studied ones: random k -SAT and random graph colouring. Our results prove that the heuristic predictions of statistical physics in this context are essentially correct. Moreover, we establish that random instances of constraint satisfaction problems have solutions well beyond the reach of any analysed algorithm.

Constraint satisfaction problems are at the heart of statistical physics, information theory and computer science. Typically, they involve a large set of variables, each taking values in a small domain, such as $\{0, 1\}$, and a collection of constraints, each binding a few of the variables by forbidding some of their possible joint values. Examples include spin-glasses in statistical physics, error-correcting codes in information theory, and satisfiability and graph colouring in computer science.

Given a collection of constraints, a fundamental scientific question is how many of them can be satisfied simultaneously. Value assignments maximizing this number are known as ground states. We are interested in 'polynomial-time' algorithms for finding ground states, that is, algorithms whose running time is bounded by a polynomial in the number of variables. In the k -SAT problem there are n binary variables $x_1, \dots, x_n$, and each constraint (k -clause) forbids precisely one out of the 2^k possible values of some $k > 2$ variables; for example, the 3-clause $x_1 \vee x_4 \vee \bar{x}_6$ means that $(x_1, x_4, x_6) \neq (0, 0, 1)$. Trivially, one can determine the ground-states of a k -SAT instance in time 2^n , but such an exhaustive search is intractable even when $n = 300$. Many problems of practical interest, such as in chip design, often have $n = 10^5$ or more variables.

Starting with Cook's pioneering work¹, since the 1970s, thousands of problems have been shown to be computationally equivalent to k -SAT, from protein-folding to aircraft-crew scheduling. That is, an efficient algorithm for k -SAT would immediately give an efficient algorithm for all such problems. It is now widely believed that no efficient algorithm exists for k -SAT, that is, that no algorithm can solve all instances efficiently. This is the famous $P \neq NP$ conjecture. At the same time, it is possible that most instances of k -SAT can be solved efficiently: perhaps, genuine hardness is only present in a minuscule fraction of all instances. As a result, a major scientific undertaking of the last twenty years has been the study of hardness in typical instances of constraint satisfaction problems (CSPs), generated by sampling uniformly at random among instances with some fixed constraints-to-variables ratio.

A breakthrough²⁻⁵ of the 1990s was the discovery that in typical

instances, hardness appears to go along with phase transitions, as suggested in the pioneering work of Fu and Anderson⁶ (for more recent accounts see also refs 7 and 8). Specifically, for many CSPs, computational experiments suggest that as constraint density increases, the probability that all constraints can be satisfied drops precipitously from near 1 to near 0; at around the same point, the complexity of finding ground-states appears to increase steeply. In the most-studied example, random instances of k -SAT are generated by sampling uniformly, independently and with replacement $m = rn$ constraints from among all possible ones on $x_1, \dots, x_n$. To understand where the really hard problems are, let us define r_k to be the largest value such that for $r < r_k$, with high probability all constraints can be satisfied. (We will say that an event occurs with high probability (w.h.p.) if its probability tends to 1 as $n \rightarrow \infty$.) (Throughout the paper, we will assume that k is arbitrarily large but fixed, while $n \rightarrow \infty$.) Similarly, define r_k^* to be the smallest value such that for $r > r_k^*$, w.h.p. not all constraints can be satisfied. The Satisfiability Threshold Conjecture asserts that, in fact, $r_k = r_k^*$ for all $k > 2$ (Fig. 1).

A very simple probabilistic counting argument implies that $r_k^* \leq 2^k \ln 2$: because constraints are chosen independently, the probability there exists at least one satisfying assignment is at most $2^n (1 - 2^{-k})^{rn}$, a quantity that tends to 0 for $r \geq 2^k \ln 2$. Heuristic techniques of statistical physics⁹⁻¹¹ also predict that the threshold scales, approximately, as $2^k \ln 2$. On the other hand, all satisfiability algorithms that have been rigorously analysed fail to find satisfying assignments for densities above $c2^k/k$ (we give the bounds corresponding to the best known c in Table 1). This creates a growing chasm between the largest density for which algorithms can provably find solutions and the smallest density for which solutions provably do not exist.

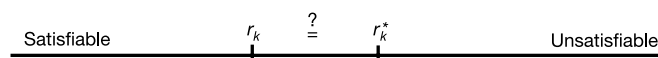


Figure 1 | The Satisfiability Threshold Conjecture. See text for details.

¹Microsoft Research, One Microsoft Way, Redmond, Washington 98052, USA. ²Department of Statistics, University of California, Berkeley, California 94720, USA.

Table 1 | Upper and lower bounds for the satisfiability threshold

k	3	4	5	7	10	20	21
Upper bound for r_k^*	4.51	10.23	21.33	87.88	708.94	726,817	1,453,635
Our lower bound for r_k	2.68	7.91	18.79	84.82	704.94	726,809	1,453,626
Algorithmic lower bound for r_k	3.52	5.54	9.63	33.23	172.65	95,263	181,453

The first row gives rigorous^{21,22} upper bounds for the location of the satisfiability threshold. The third row gives the largest densities for which some algorithm has been proved to find solutions^{23,24}. The second line represents our contribution and gives the largest densities for which we prove that solutions exist. Specifically, we prove that $r_k > 2^k \ln 2 - k$ for all k , and our lower bound converges to $2^k \ln 2 - \frac{k+1}{2} \ln 2 - 1$ as k grows.

Here we resolve this tension by proving that solutions exist much beyond the reach of all analysed algorithms. The key new element of our approach is its focus on how the structure of the space of solutions evolves as density is increased. As a result, we simultaneously get rigorous results on the location of thresholds and insights into why algorithms have such a hard time approaching them. We present the following three concrete results as an illustration of our method.

Statement of results

● We prove that for all $k \geq 3$, the satisfiability threshold lies in the interval $(2^k \ln 2 - k, 2^k \ln 2)$, thus pinpointing its location up to an exponentially small second-order term. Heuristic techniques of statistical physics predict¹² explicit values for the satisfiability threshold for each $k \geq 3$ that scales as $2^k \ln 2 - b_k$, where $b_k \rightarrow (1 + \log 2)/2$ (we discuss this point further in the ‘Discussion’ section). (See Fig. 1.)

● We rigorously determine the asymptotic threshold location for the optimization version of k -SAT, known as ‘Max k -SAT’, with exponential accuracy. Notably, for this (harder) positive-temperature problem there were no heuristic predictions using the techniques from statistical physics. (See Fig. 2.)

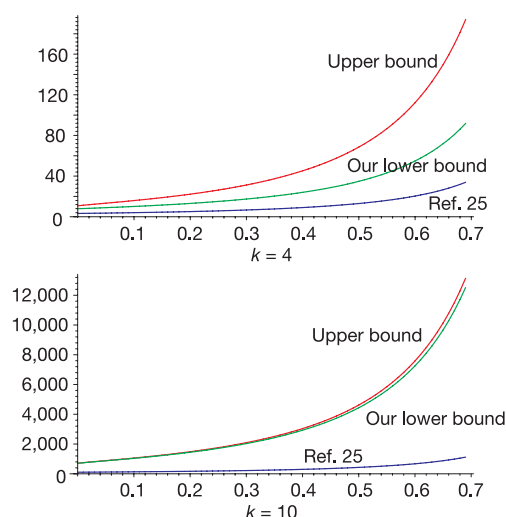


Figure 2 | Our results for random Max k -SAT. Upper and lower bounds for the critical density at which a typical k -SAT instance stops having a truth assignment that satisfies $1 - q2^{-k}$ fraction of its clauses, as a function of $q \in [0, 1)$. The points where the graphs intersect the vertical axis correspond to the bounds for the satisfiability threshold ($q = 0$). The red graph corresponds to the rigorous upper bound $T_k(q) = \lceil 2^k \ln 2 \rceil / [1 - q + q \ln q]$. The blue graph corresponds to the largest density for which any algorithm has been proved²⁵ to find assignments satisfying a $1 - q2^{-k}$ fraction of all clauses. Note that this graph is of the order of $T_k(q)/k$, thus rapidly diverging from the red graph with k . In contrast, our lower bound in the green graph converges exponentially fast to the red graph. Specifically, it is of the order of $T_k(q)(1 - \delta_k)$, where $\delta_k = O(k2^{-k/2})$.

● Given a network (graph) G , its chromatic number is the smallest number of colours with which its nodes can be coloured so that adjacent nodes receive different colours. A famous example of graph colouring is the four-colour theorem¹³, which states that any planar network (or ‘map’) has chromatic number at most four. Indeed, in our example below (Fig. 3), three colours suffice. For general, non-planar networks, though, the chromatic number can range anywhere from one to the largest degree plus one.

We prove that this variability is due to a tiny minority of networks. Specifically, we prove that if we pick a graph uniformly at random among all graphs with average degree d , then with probability that tends to 1 as n tends to infinity, its chromatic number is either k_d or $k_d + 1$, where k_d is the smallest integer k such that $d < 2k \ln k$. Thus, fixing the average degree and restricting attention to typical networks can replace planarity in yielding a chromatic number which is essentially known a priori. (See Fig. 3.)

Our method

To resolve whether the failure of algorithms was due to a genuine lack of solutions, as opposed to the difficulty of finding them, we use a method that ignores individual solutions and captures, instead, statistical properties of the entire solution-space. This statistical point of view allows us to avoid the pitfall of computational complexity; we can prove that solutions exist in random instances, without the need to identify a solution for each instance (as algorithms do).

Indeed, if random formulas are genuinely hard near the threshold, then focusing on the existence of solutions rather than their efficient discovery is essential: one cannot expect algorithms to provide accurate results on the threshold’s location; they simply cannot get there!

Our approach can be characterized as a ‘second moment’ method, as it starts from the following basic fact: every non-negative random variable X satisfies $\Pr[X > 0] \geq \mathbb{E}[X]^2 / \mathbb{E}[X^2]$. We prove the existence of solutions by applying this inequality to a random variable X that captures an appropriate weighting of the solutions in a CSP. As we will see shortly, such a weighting can be necessary. For example, in random k -SAT, if we simply let X be the number of satisfying assignments, then the ratio $\mathbb{E}[X]^2 / \mathbb{E}[X^2]$ is exponentially small in n . An important first step in mitigating this problem was made in ref. 14, where the inequality $r_k \geq 2^{k-1} \ln 2 - O(1)$ was established, by assigning non-zero weight only to those satisfying assignments whose complement is also satisfying.

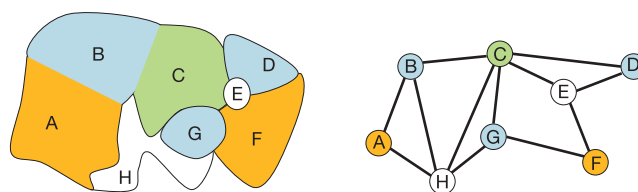


Figure 3 | A planar map and its representation as a network, both properly coloured.

The key to our approach is a systematic search for weights that asymptotically maximize the ratio $\mathbb{E}[X]^2/\mathbb{E}[X^2]$ in the class of tractable weights for which this ratio can be computed. Notably, in choosing these weights we are free to use as hints non-rigorous ideas and insights from physics, without compromising the rigour of the final result. Indeed, while physicists, mathematicians and computer scientists have been investigating the same constraint-satisfaction problems for some time now, this was mostly done using disjoint tool chests. Our approach, on the other hand, can be used in conjunction with the physics heuristics to gain further insight into the geometry of solution spaces; indeed this interaction is already taking place.

The vanilla second moment method fails on random k -SAT. Given any k -SAT instance F on n variables, let $X = X(F)$ be its number of satisfying assignments. By computing $\mathbb{E}[X^2]$ and $\mathbb{E}[X]^2$ for random formulas with a given density r , denoted as $F_k(n, rn)$, one can hope to get a lower bound on the probability that $X > 0$, that is, that $F_k(n, rn)$ is satisfiable. Unfortunately, as we show below, this direct application fails dramatically because $\mathbb{E}[X^2]$ is exponentially (in n) greater than $\mathbb{E}[X]^2$ for every density $r > 0$. Nevertheless, this estimation is useful because it points to the source of the problem and lays the foundation for our later successful choice of X .

For a k -CNF formula with independent clauses $c_1, c_2, \dots, c_m$ it is straightforward to show that:

$$\mathbb{E}[X^2] = 2^n \sum_{z=0}^n \binom{n}{z} f_S(z/n)^m \quad (1)$$

where $f_S(\alpha) = 1 - 2^{1-k} + 2^{-k}\alpha^k$ is the probability that two fixed truth assignments that agree on $z = \alpha n$ variables both satisfy a randomly drawn clause. Observe that f is an increasing function of α and that $f_S(1/2) = (1 - 2^{-k})^2$, which means that truth assignments having overlap $n/2$ are uncorrelated.

Letting $\Lambda_S(\alpha) = [2f_S(\alpha)^r]/[\alpha^\alpha(1-\alpha)^{1-\alpha}]$ we see that $\mathbb{E}[X]^2 = (2^n(1 - 2^{-k})^m)^2 = (4f_S(1/2)^m)^2 = \Lambda_S(1/2)^n$. Because $\binom{n}{\alpha n} = (\alpha^\alpha(1-\alpha)^{1-\alpha})^{-n} \times \text{poly}(n)$ we see from equation (1) that $\mathbb{E}[X^2] \geq (\max_{0 \leq \alpha \leq 1} \Lambda_S(\alpha))^n \times \text{poly}(n)$. Therefore, if there exists some $\alpha \neq 1/2$ such that $\Lambda_S(\alpha) > \Lambda_S(1/2)$, then the second moment is exponentially greater than the square of the expectation and we get only an exponentially small lower bound for $\Pr[X > 0]$. Put differently, unless the dominant contribution to $\mathbb{E}[X^2]$ comes from uncorrelated pairs of satisfying assignments, that is, pairs with overlap $n/2$, the second moment method fails. Unfortunately, for all $r > 0$, we have $\Lambda'_S(1/2) \neq 0$. This is because the entropic factor $\varepsilon(\alpha) = 1/(\alpha^\alpha(1-\alpha)^{1-\alpha})$ is symmetric around $\alpha = 1/2$, while f_S is increasing in $(0, 1)$. As a result, the derivative of Λ_S becomes 0 only when the correlation benefit balances with the penalty of decreasing entropy at some $\alpha > 1/2$. We demonstrated this, for $k = 5$, in Fig. 4.

In general, given a real number $\alpha \in [0, 1]$, we would like to know the number of pairs of satisfying truth assignments that agree on $z = \alpha n$ variables in a random formula. Each term in the sum in equation (1) gives us the expected number of such pairs. Although this expectation overemphasizes formulas with more satisfying assignments, it gives valuable information on the distribution of distances among truth assignments in a random formula. For example, if for some values of z (and k, r) this expectation tends to 0 with n , we can infer that w.h.p. there are no pairs of truth assignments that agree on z variables in a random k -CNF formula with density r .

Balance and the weighted second moment method. An attractive feature of the second moment method is that we are free to apply it to any random variable $X = X(F)$ such that $X > 0$ implies F is satisfiable. Sums of the form $X = \sum_\sigma w(\sigma, F)$ clearly have this property as long as $w(\sigma, F) = 0$ when σ is not a satisfying assignment. Such weighting schemes can be viewed as transforms of the original problem and can be particularly effective in exploiting insights into the source of correlations.

With this in mind, let us consider random variables of the form $X = \sum_\sigma \Pi_c w(\sigma, c)$, where w is some arbitrary function. (Eventually, we will require that $w(\sigma, c) = 0$ if σ falsifies c .) Similarly to equation (1), it is rather straightforward to prove that $\mathbb{E}[X^2] = 2^n \sum_{z=0}^n \binom{n}{z} f_w(z/n)^m$, where $f_w(z/n) = \mathbb{E}[w(\sigma, c)w(\tau, c)]$ is the correlation between two truth assignments σ and τ that agree on z variables, with respect to a single random clause c . It is also not hard to see that $f_w(1/2) = \mathbb{E}[w(\sigma, c)]^2$, that is, truth assignments at distance $n/2$ are uncorrelated for any function w . Thus, arguing as in the previous section, we see that $\mathbb{E}[X^2]$ is exponentially greater than $\mathbb{E}[X]^2$ unless $f'_w(1/2) = 0$.

At this point we observe that because we are interested in random formulas where literals are drawn uniformly, it suffices to consider functions w such that: for every truth assignment σ and every clause $c = \ell_1 \vee \dots \vee \ell_k$, $w(\sigma, c) = w(\mathbf{v})$, where $v_i = +1$ if ℓ_i is satisfied under σ and $v_i = -1$ if ℓ_i is not satisfied under σ . (So, we will require that $w(-1, \dots, -1) = 0$.) Letting $A = \{-1, +1\}^k$ and differentiating f_w yields the geometric condition:

$$f'_w(1/2) = 0 \Leftrightarrow \sum_{\mathbf{v} \in A} w(\mathbf{v}) \mathbf{v} = 0 \quad (2)$$

The condition in the right-hand side (r.h.s.) of equation (2) asserts that the vectors in A , when scaled by $w(\mathbf{v})$, must cancel. This gives us another perspective on the failure of the vanilla second moment method: when $w = w_S$ is the indicator variable for satisfiability, the condition in the r.h.s. of equation (2) does not hold because the

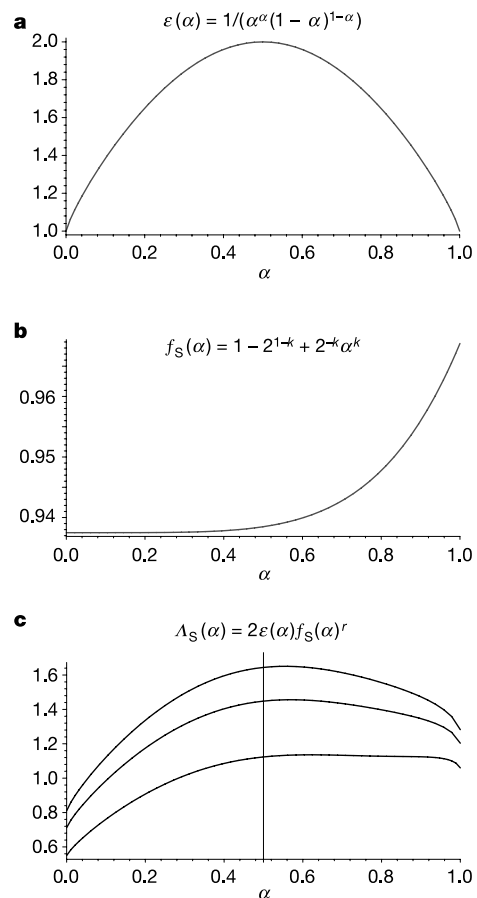


Figure 4 | Plots of entropy, correlation, and their product for the vanilla second moment method. The function f_S is plotted for $k = 5$. The function Λ_S is plotted for $k = 5$ with different values of r . Specifically, $r = 14, 16, 20$ from top to bottom. The vertical line at $\alpha = 1/2$ highlights that Λ_S is maximized for $\alpha > 1/2$.

vector $(-1, -1, \dots, -1)$ has weight 0, while all other $\mathbf{v} \in A$ have weight 1.

Note that the r.h.s. of equation (2) implies that in a successful weighting each coordinate must have mean 0: each literal must be equally likely to be +1 or -1 when we pick truth assignments with probability proportional to their weight under w . We will call truth assignments with $km/2$ satisfied literal occurrences 'balanced'.

To make the second moment method work we would like to choose a function w that is 'as close as possible' to w_S while being balanced, that is, satisfy equation (2). For $\mathbf{v} \in A$, let $|\mathbf{v}|$ denote its number of +1 values. Maximizing the relative entropy of w with respect to w_S subject to equation (2) yields:

$$w(\mathbf{v}) = \begin{cases} 0 & \text{if } \mathbf{v} = (-1, \dots, -1) \\ \lambda^{|\mathbf{v}|} & \text{otherwise} \end{cases} \quad (3)$$

where λ satisfies $(1 + \lambda)^{k-1} = 1/(1 - \lambda)$. In Fig. 5 we plot the functions f_w and Λ_w corresponding to the w in equation (3) for the same choice of k, r as in Fig. 4. For $k \geq 3$, we prove¹⁵ that Λ_w is maximized at $\alpha = 1/2$ as long as $r \leq 2^k \ln 2 - (k+1)\frac{\ln 2}{2} - 1 - \beta_k$, where $\beta_k \rightarrow 0$. This implies that for such r , $\mathbb{E}[X^2] < C \cdot \mathbb{E}[X]^2$, where $C = C(k) > 0$ is independent of n . By the second moment method this implies $\Pr[X > 0] \geq 1/C$ and by a result of Friedgut¹⁶, it follows that $r_k \geq r$.

To gain additional insight into balanced assignments it helps to think of $F_k(n, m)$ as generated in two steps: first choosing the km literal occurrences randomly, and then partitioning them into clauses. Already, at the end of the first step, truth assignments that satisfy many literal occurrences have significantly greater conditional probability of being satisfying. But such assignments are highly correlated because in order to satisfy many literal occurrences they tend to agree with the majority truth assignment, and thus each other, on more than half the variables. Our choice of λ penalizes satisfying assignments that satisfy more than half of all literal occurrences in the formula: that is, more than a random assignment. In other words, our random variable X curbs the tendency of satisfying assignments to lean towards the majority vote assignment. The fact that Λ_w is maximized at $\alpha = 1/2$ for densities almost all the

way to the random k -SAT threshold, means that for all such densities the expected number of pairs of balanced assignments at distance $n/2$ is exponentially greater than the expected number of pairs of balanced assignments at any other distance.

Random Max k -SAT

Say that a k -SAT instance with m clauses is ' p -satisfiable', where $p \in [0, 1]$, if there exists an assignment satisfying at least $(1 - 2^{-k} + p2^{-k})m$ clauses (observe that every k -CNF formula is 0-satisfiable since a random truth assignment satisfies $(1 - 2^{-k})m$ clauses in expectation). We define $r_k(p)$ to be the largest value such that for $r < r_k(p)$, w.h.p. a random formula $F_k(n, rn)$ is p -satisfiable. Similarly, we define $r_k^*(p)$ to be the smallest value such that for $r > r_k^*(p)$, w.h.p. a random formula $F_k(n, rn)$ is not p -satisfiable. So, in these terms, the Satisfiability Threshold Conjecture amounts to $r_k(1) = r_k^*(1)$.

As we saw earlier for the case $p = 1$, that is, for satisfiability, a major factor in the excessive correlations behind the failure of the vanilla second moment method is that satisfying truth assignments tend to lean toward the majority vote truth assignment. To avoid this pitfall, again, we consider balanced truth assignments, this time p -satisfying ones. Specifically, for $\sigma \in \{0, 1\}^n$ we let:

- (1) $H = H(\sigma, F)$ be the number of satisfied literal occurrences in F under σ , minus the number of unsatisfied literal occurrences in F under σ , and
- (2) $U = U(\sigma, F)$ be the number of unsatisfied clauses in F under σ , minus $m(1 - p)/2^k$.

To focus on the desired truth assignments we fix $\beta > 0$ and $0 < \gamma < 1$ and define $X(\beta, \gamma)$ as:

$$X(\beta, \gamma) = \sum_{\sigma} \gamma^{H(\sigma, F)} e^{-\beta U(\sigma, F)} \quad (4)$$

Because $\beta > 0$ and $0 < \gamma < 1$ we see that the truth assignments σ for which $H(\sigma, F) > 0$ or $U(\sigma, F) > 0$ are suppressed exponentially, while the rest are rewarded exponentially. In statistical physics terms, β can be interpreted as an inverse temperature, where perfect satisfiability, analysed in the previous section, corresponds to 0 temperature, that is, $\beta \rightarrow \infty$.

By applying the second moment method to X we pinpoint¹⁷ the values of $r_k(p)$ and $r_k^*(p)$ with relative error that tends to zero exponentially fast in k . Specifically, for every $p \in (0, 1)$ let:

$$T_k(p) = \frac{2^k \ln 2}{p + (1-p) \ln(1-p)}$$

and define $T_k(1) = 2^k \ln 2$ so that $T_k(\cdot)$ is continuous on $(0, 1]$.

Theorem 1

There exists a sequence $\delta_k = O(k2^{-k/2})$, such that for all $k \geq 2$ and $p \in (0, 1]$:

$$(1 - \delta_k)T_k(p) < r_k(p) \leq r_k^*(p) < T_k(p)$$

Random graph colouring

To generate a typical network with n nodes and average degree d , we start with n isolated nodes and join each pair of them with probability $p = d/(n-1)$, independently of all others. This is known as the Erdős-Rényi $G(n, p)$ model of random graphs. We prove:

Theorem 2

For any real number $d > 0$, let k_d be the smallest integer k such that $d < 2k \ln k$. With high probability the chromatic number of a random $G(n, p)$ graph with average degree d is either k_d or $k_d + 1$.

To prove Theorem 2 we start with n isolated nodes and repeat m times: pick two nodes uniformly, independently, with replacement, and join them. We claim that for every $d > 0$, if $m = dn/2$, then w.h.p. the chromatic number of the resulting (multi)graph G is either

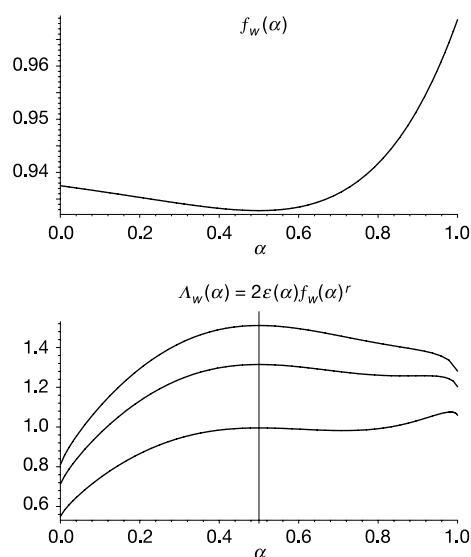


Figure 5 | Plots of the correlation function and its product with entropy for the weighted second moment method, when the weighing is given by equation (3). The function f_w is plotted for $k = 5$ with the same values of r as in Fig. 2, namely, $r = 14, 16, 20$ from top to bottom. The vertical line at $\alpha = 1/2$ shows that Λ_w is maximized at $\alpha = 1/2$ for $r = 14$ and $r = 16$, but at some $\alpha > 1/2$ for $r = 20$.

k_d or $k_d + 1$. By standard results of random graph theory¹⁸, this implies Theorem 2. To prove our claim we first observe that the probability that G is k -colourable is at most $k^n(1 - 1/k)^{dn/2}$, which tends to 0 for $d > 2k \ln k$. Our main contribution is to prove that for slightly smaller d , namely $d < 2(k - 1) \ln(k - 1)$, w.h.p. G is k -colourable. To do that we proceed as follows.

Let $X = X(G)$ be the number of balanced k -colourings of G , that is, k -colourings in which all colour classes have equal size. We apply the second moment method to X . To compute $\mathbb{E}[X^2]$ we consider all pairs σ, τ of candidate solutions (all pairs of balanced k -partitions of the n vertices) and for each such pair determine the probability that both k -partitions will assign distinct colours to the endpoints of a randomly drawn edge. If a_{ij} is the number of vertices having colour i in σ and colour j in τ it is not hard to see that this probability is:

$$1 - \frac{2}{k} + \sum_{i,j} a_{ij}^2 \quad (5)$$

Equation (5) highlights the main difficulty we need to overcome in this context: whereas in satisfiability problems the overlap parameter for a pair of solutions was a single integer (their Hamming distance), now it is a $k \times k$ matrix. As a result, to determine $\mathbb{E}[X^2]$ we need to resolve an entropy–energy tradeoff over doubly stochastic matrices, a problem of major analytic difficulty.

More precisely, we show that $\mathbb{E}[X^2]$ is dominated by the contribution of the pairs of k -partitions whose overlap matrix A maximizes the function:

$$g_d(A) = - \sum_{i,j} a_{ij} \log a_{ij} + \frac{d}{2} \cdot \log \left(1 - \frac{2}{k} + \sum_{i,j} a_{ij}^2 \right)$$

The first term in g_d measures the number of pairs of k -partitions that have A as their overlap matrix, while the second term measures the probability that such pairs will be valid k -colourings in a random graph with average degree d . When $a_{ij} = 1/k^2$ for all i, j , corresponding to uncorrelated partitions, the first term is maximized while the second is minimized. At the other extreme, when each row and column has precisely one element equal to $1/k$, corresponding to perfectly correlated partitions, the first term is minimized while the second term is maximized. If we interpret A as spreading a fixed amount of mass over k^2 cells, we see that there are two ‘forces’ that determine its shape: entropy, favouring flatness, and energy, favouring the formation of peaks.

To resolve this entropy–energy tradeoff we develop general optimization tools that we expect to be of much broader applicability. In particular, we prove that as d is increased the maximizer switches instantaneously from the perfectly flat matrix J_k to a matrix in which more than half the mass is captured by only k entries. At that point the dominant contribution to $\mathbb{E}[X^2]$ stops corresponding to uncorrelated k -colourings and the second moment fails. Our result follows by proving¹⁹ that the switch in the locus of the maximizer occurs for some $d > 2(k - 1) \ln(k - 1)$.

Discussion

Our work implies that the physics predictions for certain key combinatorial optimization and decision problems can be rigorously justified. Indeed, recently, non-rigorous techniques of statistical physics were used to derive¹¹ predictions for the location of the threshold for random k -SAT for every fixed $k \geq 3$; for example, for $k = 3, 4, 5$ the predicted values are 4.267, 9.931 and 21.117, respectively. These predictions are compatible with the rigorous bounds and indeed, match the rigorous upper bound as $k \rightarrow \infty$. Similarly, for the graph colouring problem it was predicted in ref. 20 that the k -colourability transition occurs at $d = 2k \log k - 1 + o(1)$, which fits neatly between the rigorous upper and lower bounds we state in the ‘Random graph colouring’ section.

The gap between our lower bound for the location of the

k -SAT threshold and the best-known algorithmic lower bound $r_k = \Omega(2^k/k)$, seems to us the most significant remaining problem. Indeed, the physics ideas mentioned above have motivated a new satisfiability algorithm^{10,11} that performs extremely well for small values of k , such as $k = 3$. A rigorous analysis of this algorithm is still lacking, and it remains unclear whether its success for values of r close to the threshold extends to large k . Indeed, even evaluating the algorithm experimentally is already intractable for moderate values of k , such as $k = 10$.

The success of the second moment method for balanced satisfying truth assignments suggests that such assignments form a ‘mist’ in $\{0, 1\}^n$ and, as a result, they might be hard to find by algorithms based on local updates. (More precisely, the satisfying assignments decompose into clusters whose diameter is much smaller than the inter-cluster distances, as predicted in ref. 11.) Moreover, as k increases the influence exerted by the majority vote assignment becomes less and less significant as most literals occur very close to their expected $kr/2$ times. As a result, as k increases, typical satisfying assignments get closer and closer to being balanced, meaning that the structure of the space of solutions for small values of k might be significantly different from the structure for large values of k . Indeed, an appealing intuitive explanation for the fact that our methods succeed where all analysed algorithms fail is that the latter rely on knowing part of a solution to get an indication on the values of the unknown variables. The success of such a strategy inevitably requires that the space of solutions is clustered, which means that solutions are highly correlated. Our method, on the other hand, targets situations in which solutions form a ‘sparse mist’ in the configuration space. Thus, it draws strength precisely from the phenomenon that causes algorithms to fail.

To summarize, the following key question remains: is there an algorithmic threshold $\lambda_k = o(2^k)$ so that for densities $r > \lambda_k$, no polynomial-time algorithm can find a satisfying truth assignment with probability bounded away from 0 as $n \rightarrow \infty$?

Received 16 December 2004; accepted 31 March 2005.

1. Cook, S. A. The complexity of theorem-proving procedures. *Proc. 3rd Ann. ACM Symp. on Theory of Computing* 151–158 (1971).
2. Cheeseman, P., Kanefsky, B. & Taylor, W. Where the really hard problems are. *Proc. 12th Int. Joint Conf. on Artificial Intelligence* 331–337 (1991).
3. Selman, B., Levesque, H. & Mitchell, D. Hard and easy distributions of SAT problems. *Proc. 10th Nat. Conf. on Artificial Intelligence* 459–465 (1992).
4. Kirkpatrick, S. & Selman, B. Critical behavior in the satisfiability of random boolean expressions. *Science* **264**, 1297–1301 (1994).
5. Monasson, R., Zecchina, R., Kirkpatrick, S., Selman, B. & Troyansky, L. Determining computational complexity from characteristic ‘phase transitions’. *Nature* **400**, 133–137 (1999).
6. Fu, Y. & Anderson, P. W. Application of statistical mechanics to NP-complete problems in combinatorial optimisation. *J. Phys. A* **19**, 1605–1620 (1986).
7. Anderson, P. W. Solving problems in finite time. *Nature* **400**, 115–116 (1999).
8. Gomes, C. P. & Selman, B. Satisfied with physics. *Science* **297**, 784–785 (2002).
9. Monasson, R. & Zecchina, R. Statistical mechanics of the random K -satisfiability model. *Phys. Rev. E* **56**, 1357–1370 (1997).
10. Mézard, M. & Zecchina, R. Random K -satisfiability: from an analytic solution to a new efficient algorithm. *Phys. Rev. E* **66**, 056126 (2002).
11. Mézard, M., Parisi, G. & Zecchina, R. Analytic and algorithmic solution of random satisfiability problems. *Science* **297**, 812–815 (2002).
12. Mertens, S., Mézard, M. & Zecchina, R. Threshold values of random k -SAT from the cavity method. To appear in *Random Struct. Algorithms* (in the press); preprint at <http://arxiv.org/abs/cs.CC/0309020> (2003).
13. Appel, K. & Haken, W. Every planar map is four colorable. *Contemp. Math.* **98**, (1989).
14. Achlioptas, D. & Moore, C. The asymptotic order of the random k -SAT threshold. *Proc. 43rd Ann. IEEE Symp. on Foundations of Computer Science* 779–788 (2002).
15. Achlioptas, D. & Peres, Y. The threshold for random k -SAT is $2^k \log 2 - O(k)$. *J. Am. Math. Soc.* **17**, 947–973 (2004).
16. Friedgut, E. Sharp thresholds of graph properties, and the k -SAT problem. *J. Am. Math. Soc.* **12**, 1017–1054 (1999).
17. Achlioptas, D., Naor, A. & Peres, Y. On the maximum satisfiability of random formulas. *Proc. 44th Ann. IEEE Symp. on Foundations of Computer Science* 362–370 (2003).

18. Bollobás, B. *Random Graphs* Vol. 73, 2nd edn *Cambridge Studies in Advanced Mathematics* (Cambridge Univ. Press, Cambridge, 2001).
19. Achlioptas, D. & Naor, A. The two possible values of the chromatic number of a random graph. *Proc. 36th Ann. ACM Symp. on Theory of Computing* 587–593 (2004).
20. Krzakala, F., Pagnani, A. & Weigt, M. Threshold values, stability analysis and high-q asymptotics for the coloring problem on random graphs. *Phys. Rev. E* **70**, 046705 (2004).
21. Kirousis, L. M., Kranakis, E., Krizanc, D. & Stamatiou, Y. Approximating the unsatisfiability threshold of random formulas. *Random Struct. Algorithms* **12**, 253–269 (1998).
22. Dubois, O., Boufkhad, Y. & Mandler, J. Typical random 3-SAT formulae and the satisfiability threshold. *Proc. 11th Ann. ACM-SIAM Symp. on Discrete Algorithms* 126–127 (2000).
23. Kaporis, A. C., Kirousis, L. M. & Lalas, E. G. The probabilistic analysis of a greedy satisfiability algorithm. *Proc. 10th Ann. European Symp. on Algorithms* 574–585 (2002).
24. Frieze, A. M. & Suen, S. Analysis of two simple heuristics on a random instance of k -SAT. *J. Algorithms* **20**, 312–355 (1996).
25. Coppersmith, D., Gamarnik, D., Hajiaghayi, M. T. & Sorkin, G. B. Random MAX SAT random MAX CUT, and their phase transitions. *Random Struct. Algorithms* **24**, 502–545 (2004).

Acknowledgements We thank S. Kirkpatrick, J. Kleinberg and S. Mertens for feedback on the presentation of the results.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.A. (optas@microsoft.com).

Structural insights into a yeast prion illuminate nucleation and strain diversity

Rajaraman Krishnan¹ & Susan L. Lindquist¹

Self-perpetuating changes in the conformations of amyloidogenic proteins play vital roles in normal biology and disease. Despite intense research, the architecture and conformational conversion of amyloids remain poorly understood. Amyloid conformers of Sup35 are the molecular embodiment of the yeast prion known as [PSI], which produces heritable changes in phenotype through self-perpetuating changes in protein folding. Here we determine the nature of Sup35's cooperatively folded amyloid core, and use this information to investigate central questions in prion biology. Specific segments of the amyloid core form intermolecular contacts in a 'Head-to-Head', 'Tail-to-Tail' fashion, but the 'Central Core' is sequestered through intramolecular contacts. The Head acquires productive interactions first, and these nucleate assembly. Variations in the length of the amyloid core and the nature of intermolecular interfaces form the structural basis of distinct prion 'strains', which produce variant phenotypes *in vivo*. These findings resolve several problems in yeast prion biology and have broad implications for other amyloids.

Prion proteins share an unusual property: they adopt distinct functional and conformational states that self-perpetuate through protein-conformational chain reactions^{1–4}. The first known prion, PrP, facilitates the transmission of a fatal neurodegenerative disease in mammals (spongiform encephalopathy) by converting non-prion conformers to the prion state^{4,5}. Fungal prions, however, are not generally pathogenic. Instead, they act as protein-only elements of inheritance. Their prion conformers produce new phenotypes—often beneficial phenotypes³—by changing processes as diverse as translation termination, nitrogen metabolism and heterokaryon formation⁶. These phenotypes are heritable because mother cells pass prion conformers on to their daughters, perpetuating the cycle of conversion¹. In the marine snail *Aplysia*, a neuronal form of CPEB, a protein implicated in long-term memory⁷, can also switch to a self-perpetuating prion conformation. In this case, conformational switching activates the protein, suggesting that CPEB's self-perpetuating prion conformation functions in the long-term maintenance of synapses⁸. Most proteins can form amyloids under specific, unusual conditions⁹. Prion proteins access such states under normal physiological conditions, and some have been conserved for hundreds of millions of years¹⁰. Prions play a much broader role in biology than initially suspected^{11–13}.

To understand these self-perpetuating conformational changes, we investigated [PSI⁺], a highly conserved prion in *Saccharomyces cerevisiae*. [PSI⁺] confers a wide variety of novel phenotypes by facilitating the read-through of nonsense codons^{11–13}. Read-through occurs when Sup35, a translation termination factor, is inactivated by conversion to an amyloid with a self-sustaining structure^{1,2,6}.

Sup35 has three distinct regions^{14,15}: C, a GTP-binding domain at the carboxy terminus; M, a highly charged middle region; and N, a glutamine/asparagine-rich amino-terminal region containing oligopeptide repeats. C facilitates translation termination, and N and M govern prion status. N is essential for converting Sup35 to the prion state *in vivo*¹⁶ and for converting soluble protein into amyloid fibres *in vitro*¹⁷. M confers solubility in the non-prion state and stabilizes the prion during mitosis and meiosis¹⁸. When N and M are removed from C and fused to the glucocorticoid receptor, they create

a new prion that confers a hormone-response phenotype on yeast but otherwise recapitulates all of the unusual physical and genetic behaviours of [PSI⁺]¹⁹. Thus, NM encodes prion function.

In vitro, the conversion of natively unfolded NM to β -rich amyloid fibres involves (1) a lag phase, in which part of the protein oligomerizes and converts to an amyloidogenic nucleus, and (2) an assembly phase, in which soluble proteins rapidly associate with mature nuclei and convert to amyloid^{17,20–24}. NM can form distinct types of self-perpetuating conformers^{17,25}. These produce distinct phenotypes (prion strains or variants) when used to transform non-prion [*psi*[−]] cells to the prion [PSI⁺] state^{26,27}. Thus, NM amyloids also embody prion structure.

The β -strands in NM fibres run perpendicular to the main axis and are spaced ~ 4.7 Å apart^{20,28}. Several mutations in NM are known to affect prion maintenance and fibre assembly^{17,18,29,30}. However, we do not yet know (1) the arrangement of individual NM molecules within the amyloid fibre, (2) the mechanisms of nucleation and conformational conversion, or (3) the structural basis of prion strains. To address these questions, we capitalized on the absence of cysteine residues in NM to create a large number of variants, each containing a single cysteine at a different location, which could be modified with fluorophores and crosslinkers.

Cysteine variants behave like wild-type NM

We created 37 individual cysteine-substitution mutations throughout NM (Fig. 1a). Each was used to replace the NM portion of the wild-type SUP35 gene *in vivo*, resulting in a single, full-length functional SUP35 gene. All NM variants retained the capacity to support the [PSI⁺] and [*psi*[−]] states (Supplementary Fig. S1a). None altered the stability of those states (data not shown).

Next, each protein was expressed in and purified from *Escherichia coli*. All spontaneously assembled into amyloid at the same rate as wild-type (Supplementary Fig. S1b). Moreover, all fibres were indistinguishable from wild-type by electron microscopy and SDS (sodium dodecyl sulphate) solubility²² (data not shown). Finally, fibres made from each mutant seeded assembly as well as the wild type^{17,20} (Supplementary Fig. S1c). Having established that

¹Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA.

cysteine-substitution proteins recapitulate prion behaviour both *in vivo* and *in vitro*, we used them to explore NM structure and assembly.

Boundaries of the cooperatively folded amyloid

We used two independent approaches to investigate the general structure of NM fibres. First, in 37 sets of fibres assembled at 25 °C from each variant, we probed the accessibility of cysteine residues to labelling with pyrene maleimide, and to a more hydrophilic reagent, Lucifer yellow (Fig. 1b and data not shown). Proteins carrying a cysteine residue between amino acid residues 25 and 58 were sparsely labelled with either pyrene maleimide or Lucifer yellow. Proteins with a cysteine residue in amino acids 2–21 and 68–112 showed partial accessibility. All cysteines in M were highly accessible.

Next, we used acrylodan to report on the conformational status of different segments of NM. Acrylodan shows an increase in fluorescence intensity and a blue shift in $\lambda_{\max}$ when sequestered from

solvent. When denatured, all cysteine-labelled proteins had a $\lambda_{\max}$ around 530 nm. After assembly at 25 °C, proteins labelled with acrylodan at a cysteine between residues 21 and 121 had strongly blue-shifted emissions ($\lambda_{\max}$ 486–488 nm), indicating sequestration from solvent (no guanidine hydrochloride (GdmCl) in Fig. 1c and data not shown). Proteins labelled in regions adjoining residues 21–121 (residue 7 N-terminally and residues 137, 158 and 167 C-terminally) had partially blue-shifted emission maxima ($\lambda_{\max}$ 493–525 nm). Proteins labelled at residues 2, 184, 225 or 234 had no significant blue shift, indicating that these residues remained exposed to solvent pre- and post-assembly (Fig. 1c and data not shown).

To determine which residues participate in the same cooperatively folded structure, we assessed their post-assembly GdmCl denaturation profiles, using 24 different GdmCl concentrations for each of the fibres. All fibres labelled between residues 21 and 121 showed a similar drop in fluorescence intensity and a corresponding red shift in emission maxima, with an inflection at $2.5 \text{ M} \pm 0.15 \text{ M}$ GdmCl (Fig. 1c). These profiles fitted a monophasic unfolding transition that corresponds to the major unfolding transition of wild-type NM protein (A. Cashikar and S. L. L., unpublished results). Adjacent cysteine variants that had intermediate blue shifts upon fibrillization also had distinct unfolding transitions in GdmCl.

These studies establish that (1) a large portion of N is sequestered from solvent; (2) a sub-portion of N constitutes a distinct domain (formed by contiguous amino acids, including 21–121) with an unusually stable structure and a single cooperative unfolding transition; (3) flanking sequences are structurally heterogeneous, with residues 137 and 158 having distinct, but cooperative, unfolding transitions, and residues 2 and 7 having multiphasic transitions; and (4) M is flexible and solvent exposed from residue 158 onwards. Residues in the N/M transition zone (amino acids 121, 137 and 158) were fully accessible to post-assembly cysteine labelling but according to acrylodan fluorescence and GdmCl denaturation were partially sequestered and structured. This probably reflects different sensitivities of the two techniques used to detect structural instability. For example, if residue 121 occasionally adopts an open structure, it would show a strong blue shift with acrylodan but still be accessible to prolonged labelling.

Identifying intermolecular contacts

To determine which regions of NM make intermolecular contacts in fibres, we exploited the ability of pyrene-labelled proteins to form excimers (excited-state dimers). When two pyrene molecules lie within 4–10 Å of each other, the long fluorescence lifetime of pyrene allows an excited residue to interact with an unexcited residue before energy is emitted. This produces a strong red shift in fluorescence. Excimer fluorescence will occur between residues at β -strands that form the interface between two monomers, but not between residues that are distant from the interface. One caveat is that the pyrenes might alter the structure formed, but control experiments eliminated this concern (see Supplementary Fig. S2c).

In denaturant, each of the pyrene-labelled proteins had multiple emission maxima between 384 and 405 nm (Supplementary Fig. S2a, b). After assembly at 25 °C, most fibres labelled in the N region showed a blue shift in fluorescence (Supplementary Fig. S2b), indicating that they were sequestered from intermolecular contacts. Six different fibre preparations individually labelled at a residue in one of two distinct regions (residues 25, 31 or 38, or residues 91, 96 or 106) produced strong red-shifted fluorescence ($\lambda_{\max} \sim 465 \text{ nm}$; see Fig. 2a and Supplementary Fig. S2a). These residues must lie at or near a contact between two NM molecules. We will refer to these intermolecular contact regions as the 'Head' (residues 25–38) and 'Tail' (residues 91–106). Residues between them, the 'Central Core' (43–85), are also part of the cooperatively folded amyloid but are sequestered from intermolecular contacts.

Pyrene fluorescence patterns were virtually identical in repeat

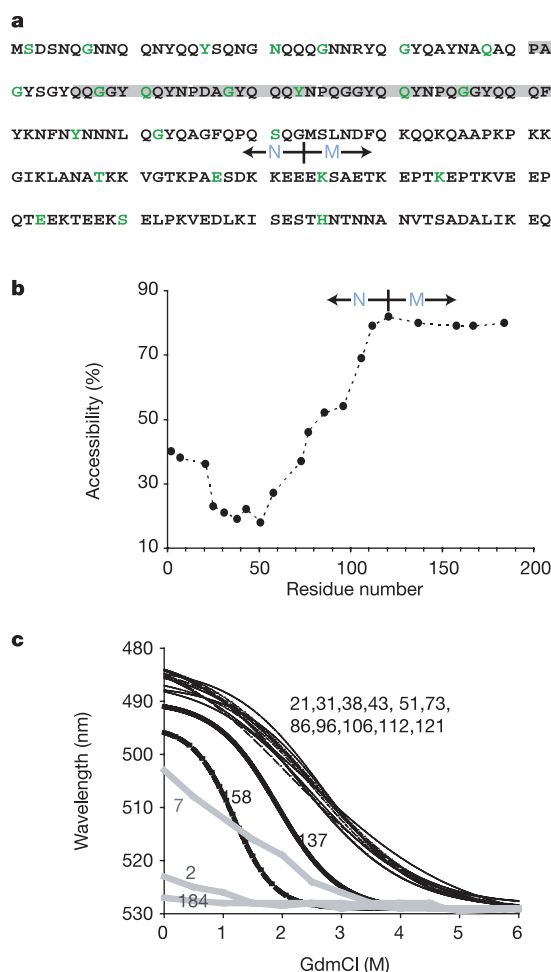


Figure 1 | Mapping the amyloid region of NM fibres. **a**, NM sequence, showing residues mutated to cysteine in green, the five imperfect oligopeptide repeats shaded in grey, and the boundary between N and M (123/124) where it is traditionally drawn. On the basis of amino acid composition, the boundary might equally be drawn at amino acids 137–139. **b**, Accessibility of cysteine residues after assembly. Fibres formed from single-cysteine substitution mutants were labelled with pyrene maleimide for 3 h. The ratio of labelled-to-unlabelled protein is plotted as a measure of accessibility. For all proteins, labelling approached 100% in denaturant (6 M GdmCl; data not shown). **c**, Denaturation profiles. Mutants were individually assembled at 25 °C, using 75% unlabelled and 25% acrylodan-labelled protein. Each preparation was then equilibrated for 30 min at 24 different concentrations of GdmCl and fluorescence spectra were recorded.

experiments (data not shown). They were also nearly identical in seeded and unseeded reactions (grey squares and black circles, respectively; Fig. 2a). The remarkable reproducibility of these excimer patterns establishes that the intermolecular contacts in seeded assembly quite precisely recapitulate those of spontaneous assembly.

Next, we mixed two proteins, each individually labelled at one cysteine residue, in all possible pair-wise combinations (Fig. 2b and Supplementary Fig. S3). Confirming that residues in the Central Core do not contribute to intermolecular contacts, all fibres in which either one or both of the proteins were labelled in this region produced low excimer signals (residues 51, 58 and 73; see Fig. 2b and Supplementary Fig. S3). The strongest signals were again from the Head and Tail regions, but only when both proteins were labelled in the same region (Fig. 2b, orange boxes). Fibres with one protein labelled in the Head and another in the Tail produced weak signals. We conclude that contacts between monomers in the fibre occur in a Head-to-Head and Tail-to-Tail fashion.

These data are not compatible with the parallel, super-pleated sheet model for the Sup35 prion domain³¹, in which individual NM molecules fold into long serpentine arrays, stacked in parallel along their entire length. They are, however, compatible with a β -helix model²⁸ and other models in which a contiguous stretch of amino acids forms a cooperative amyloid fold, and a Central Core within this region is sequestered from intermolecular contacts. Furthermore, individual subunits must form contacts in a Head-to-Head and Tail-to-Tail fashion (see Fig. 3).

Constraints on inter-subunit relationships

To provide an independent assessment of inter-subunit interactions, we introduced two types of crosslinks into each of the individual cysteine-substituted proteins under denaturing conditions. Reaction with oxidized dithiothreitol (DTT) produced disulphide bonds with a bond length of ~ 2 Å. Reaction with 1,4-bis-maleimidobutane (BMB), a homobifunctional agent, produced crosslinks with a 10.9-Å flexible linker.

Disulphide crosslinks inhibited fibre formation at every position tested between residues 21 and 121 (black bars in Fig. 2c; thioflavin T fluorescence levels equivalent to those of BSA aggregates). Disulfides in the extreme N terminus of N and in M had little effect on assembly.

In contrast, with the flexible BMB linker, NM molecules crosslinked at the Head or Tail formed fibres very efficiently (grey bars in Fig. 2c). BMB crosslinks severely impeded fibre formation only in the Central Core.

The ~ 2 -Å bond length of a disulphide bond is closer than the inter-strand distances of ~ 4.7 Å that characterize NM fibres^{20,28}. The distributions of residues for which disulphide bonds inhibit amyloid formation support our earlier conclusion that a contiguous linear segment of amino acids, including residues 21–121, constitute a cooperatively folded unit. Improper intersubunit alignments of any two residues in this region prevent folding of the rest of the domain. Residues in the extreme N terminus of N and in M are outside this domain and have little influence on its capacity to fold.

With the longer linker, apposition of two NM proteins in the Head or of two NM proteins in the Tail permits fibre formation, but apposition of Central Core residues prevents it. Thus, the separation of Central Core regions is not only a general characteristic of NM fibres but is essential for fibre formation.

We next asked whether the structural information and the tools we had assembled could be used to address two of the most enigmatic questions in prion biology. How is prion assembly nucleated? And what is the structural basis of distinct prion strains?

Early events during nucleation and assembly

First, we monitored kinetic changes in the fluorescence of proteins labelled with acrylodan. All tested proteins that were labelled in the cooperatively folded amyloid region (21–106) showed a very rapid increase in fluorescence, characteristic of a first-order reaction with no lag phase (Fig. 4a). This fluorescence increase preceded conversion to amyloid: when amyloid formation was monitored by the acquisition of an SDS-insoluble state, each acrylodan-labelled protein had the same lag and assembly phase as wild-type protein (Supplementary Fig. S4). In contrast, molecules labelled at residues 158 or 167 changed fluorescence simultaneously with amyloid formation (Fig. 4a and data not shown). Proteins labelled at residue 184, 203 or 225 showed no change in fluorescence (Fig. 4a and data not shown). We conclude that: (1) residues that form the cooperatively folded amyloid core rapidly enter a collapsed but non-amyloid state, (2) M residues proximal to N become structured only when N residues convert to amyloid, and (3) the distal region of M

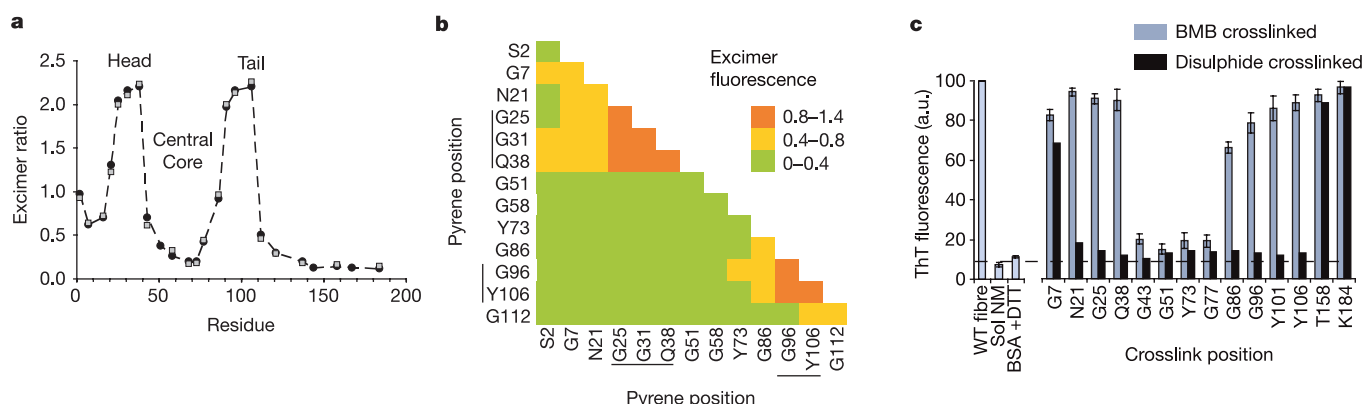


Figure 2 | Intermolecular contacts in NM fibres. **a**, Proximity analysis assessed by excimer fluorescence in fibres carrying pyrene labels at single sites. Fibres were assembled at 25 °C with 2.5 μ M NM (60% labelled), either by gently rotating the samples (black circles), or by adding 4% sonicated seed (grey squares) formed at 25 °C. The ratio of excimer fluorescence to non-excimer fluorescence (I_{465}/I_{375}) is plotted. **b**, Excimer fluorescence in fibres assembled from mixtures of two different pyrene-labelled cysteine variants. Fibres were assembled at 25 °C with gentle rotation, in reactions containing equimolar mixtures of each variant (25% labelled and 75% unlabelled). Actual excimer values are provided in Supplementary Fig. S3.

Head and Tail region residues are marked with lines. **c**, Effects of crosslinks on amyloid assembly. Proteins crosslinked under denaturing conditions with either a 2-Å disulphide bond (black bars) or a flexible 11-Å BMB crosslink (grey bars) were diluted 200-fold into buffer, to a final concentration of 2.5 μ M NM, and assembled with rotation at 25 °C. Thioflavin T (ThT) fluorescence values for wild-type fibres, soluble NM, and a denatured standard (bovine serum albumin (BSA) treated with DTT to induce denaturation) are also shown. Values represent mean $\pm$ s.d. ($n = 3$ experiments).

remains largely unstructured and exposed to solvent after amyloid assembly.

To determine which segments of the cooperatively folded amyloid region are the first to undergo conformational commitment, we took advantage of the fact that disulphide bonds anywhere in this region prohibit assembly (Fig. 2c). We reasoned that segments of NM that are the first to assume productive spatial relationships would also be the first to be protected from the spontaneous formation of disulphide bonds. Representative cysteine mutants were allowed to assemble in buffer without DTT (to facilitate the formation of disulphide bonds) and later analysed for the formation of crosslinks (dimers) on SDS gels without DTT. Cysteines in the Head region (21, 25 and 31) formed fewer disulphide bonds than cysteines at other positions (Fig. 4b). Thus, during conformational conversion, strand spacings compatible with a productive fold (and incompatible with disulphide formation) are achieved in the Head region more rapidly than in other regions.

Next, we investigated the effects of adding a single charge at various positions in the amyloid region. In some β -structures, such as the β -helix, alternating residues point towards or away from solvent, and the structures formed are very stable³². Although the introduction of a single charge would be unlikely to perturb such structures once they form, charge repulsions in early nucleating segments of the molten, collapsed intermediate would reduce the frequency with which these segments come into proximity, thereby slowing nucleation. Labelling individual NM cysteine residues with uncharged iodoacetamide had little effect on the quantity of protein converting to amyloid (Fig. 4c) or the kinetics of assembly (Supplementary Fig. S5). In contrast, iodoacetate labelling, which introduces a negatively charged moiety similar in size to acetamide, strongly inhibited assembly in the Head region (Fig. 4c and Supplementary Fig. S5).

Finally, if Head-to-Head interactions are not only characteristic of early productive amyloid conformations, but actually cause a commitment to it, bringing Head regions in proximity with each other should promote nucleation. We compared assembly kinetics for several NM variants that had been crosslinked with BMB under

denaturing conditions and then transferred to assembly buffer. Crosslinks in the Central Core (residues 43 and 73) blocked assembly entirely, confirming that these regions must be separated from each other to form amyloid (Fig. 4d). Crosslinks in the Tail (residues 96 and 106) had little effect. Crosslinks in the Head (residues 21, 25 and 38) virtually eliminated the lag phase. Thus, the juxtaposition of residues in the Head region is an early event in amyloid formation and is, indeed, sufficient to nucleate it.

Structural distinctions between prion strains

To address the second critical question in prion biology—the basis of prion ‘strains’ or variants—within the structural framework we generated for NM fibres, we assembled the protein under conditions previously known to produce fibres enriched in different different strains³² (25 °C versus 4 °C). We confirmed that fibres produced at 4 °C assembled much more rapidly²⁶ (data not shown), were less resistant to GdmCl denaturation (Fig. 5a), and produced mostly strong prion strains when used to transform cells from the [*psi*[−]] non-prion state to the [*PSI*⁺] prion state³² (Fig. 5c, left and Supplementary Fig. S7).

Do fibres enriched in different prion strains have distinct cooperatively folded amyloid domains? Denaturation profiles of fibres, independently assembled from 16 acrylodan-labelled proteins at either 4 °C or 25 °C, were determined as in Fig. 1c. In fibres at 4 °C, residues 31–86 had strong blue shifts in fluorescence upon assembly and showed a single cooperative unfolding transition at $D_{1/2} \sim 1.5$ M GdmCl (Fig. 5a and data not shown). Flanking residues 21, 25, 96, 112 and 121 had smaller blue shifts in fluorescence upon assembly, and heterogeneous denaturation profiles after assembly, as had residues flanking the amyloid domain at 25 °C (2 and 7, 137 and 158; Figs 1c and 5a, and data not shown). Thus, the cooperatively folded amyloid core has a similar character at both temperatures: it is formed by a contiguous stretch of amino acids and is flanked by residues that are structurally heterogeneous. However, the length of the region incorporated into the cooperative amyloid fold is much shorter in fibres at 4 °C than in fibres at 25 °C, and is consequently more easily denatured. Shorter, less stable amyloid cores probably

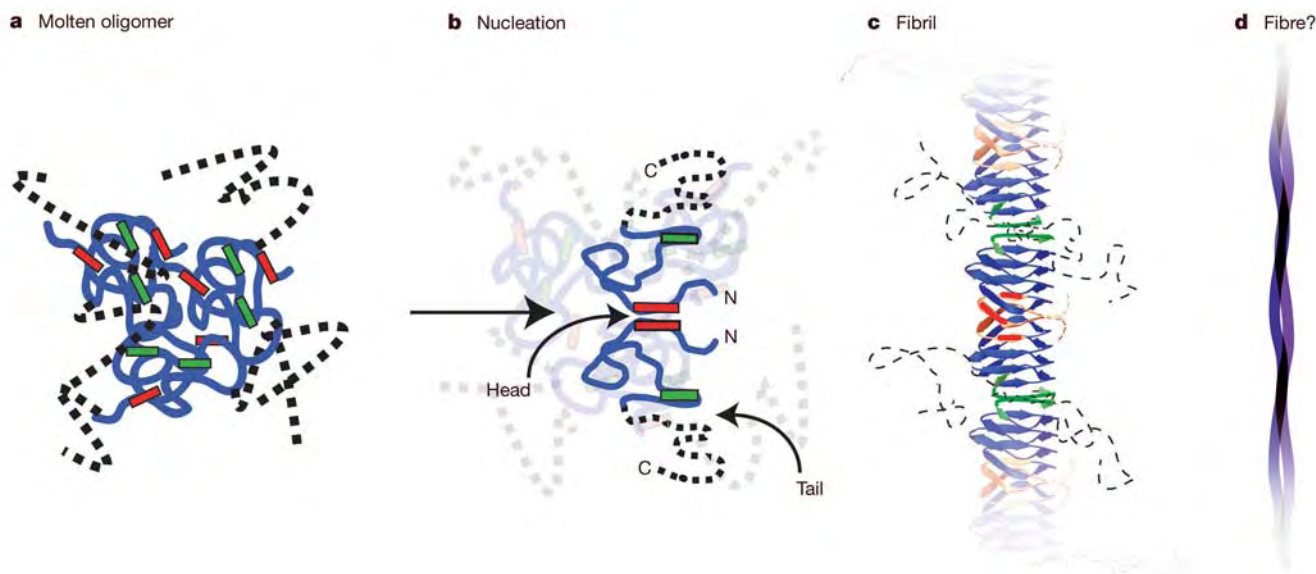


Figure 3 | An example model of NM assembly that conforms to our data. **a–d**, One model for NM assembly is provided to illustrate the constraints that our data place on the nature of NM fibre structures. In the cooperatively folded amyloid core, Head residues (red) in one NM molecule are in close proximity to Head residues of their neighbours; the same is true for the Tail residues (green) (**c**). Central Core residues (blue) are sequestered from

intermolecular interactions. M (dashed line) is largely unstructured, but the segment of M proximal to N becomes structured when amyloid forms. In the initial stages, NM molecules rapidly acquire a collapsed state (**a**) that retains a molten character until the Head regions of two molecules (**b**) come into proximity with each other and nucleate assembly. We do not know how fibrils are arranged within fibres (**d**).

produce stronger, more stably inherited prion strains because the fibres are more easily fragmented and transmitted to daughter cells, promoting the cycle of inheritance.

Finally, we asked if intermolecular contacts differ in fibres assembled at 4 °C and 25 °C. The excimer fluorescence of residues in the Head region and in flanking residues (2, 7 and 16) changed modestly but in an extremely reproducible manner (seeded and unseeded reactions; Fig. 5b). Excimer fluorescence in the Tail showed a larger shift. Thus, as the number of residues constituting the cooperatively folded amyloid domain changes, the intersubunit interfaces change as well.

The structural basis of prion strains

In fibres enriched for distinct prion strains, we confirmed²⁷ differences in the rate at which assembly occurred, and discovered differences in both the length of the amyloid core and the intersubunit contacts. But are these features determinative? If so, proteins crosslinked in different places should strongly bias assembly reactions towards different prion strains.

Proteins crosslinked with BMB (as for Fig. 2c) were assembled at 25 °C or 4 °C and used to transform cells from the [*psi*[−]] to the [*psi*⁺] state^{26,27}. Proteins crosslinked in the Central Core did not

induce [*psi*⁺] above background levels (Fig. 5c), consistent with their failure to undergo amyloid assembly (Fig. 2c). Crosslinks in the Head, which caused very rapid assembly, biased fibres towards the production of strong strains. Conversely, crosslinks in the Tail, which should cause a longer segment to be incorporated into the cooperative amyloid fold, biased fibres towards weak strains (Fig. 5c). Notably, the strain biases produced by different crosslink positions overcame those of different assembly temperatures. Whether the fibres used for transformation had been assembled at 4 °C or at 25 °C, proteins crosslinked in the Head produced primarily strong strains, and proteins crosslinked in the Tail produced primarily weak strains (Fig. 5d). We conclude that strain distinctions are due to differences in the length of the amyloid core of individual NM molecules, as well as the nature of NM–NM interfaces.

Discussion

Our results provide a framework for the structure of NM fibres, define rate-limiting events that govern their nucleation and establish a physical basis for prion strains. The data provide many insights regarding prion initiation and propagation, and are likely to shed light on amyloid biology in other systems.

The first step in assembly is the formation of a collapsed intermediate, in which residues in the N region are sequestered from solvent but have not yet locked into their final structures (Fig. 4a and Supplementary Fig. S4). The search for structure in globular proteins also begins with a collapsed intermediate, but these are dominated by hydrophobic interactions. The collapse of N must be governed by polar interactions and/or backbone interactions, because polar residues outweigh hydrophobic residues by 16 to 1. Hence, proteins governed by very different interactions, and achieving very different final structures, both do so through a molten, collapsed intermediate.

Within this collapsed state, it is the Head region that commits most rapidly to amyloid-compatible spatial relationships (Fig. 4b). Moreover, simply bringing the Head regions of two NM molecules into close proximity triggers nucleation (Fig. 4d). We suggest that Head-to-Head contacts are a principal route for NM nucleation, funneling the protein away from other folding pathways that might still be productive but are much slower. The lag phase of assembly would then represent the time required to search out these productive contacts in the collapsed molten state. The importance of Head-to-Head contacts in nucleation clarifies why the 'species barrier', which prevents cross-species seeding between *S. cerevisiae* and *Candida albicans* NM proteins, maps to this region^{2,33,34}.

After nucleation, fibres grow through Head-to-Head and Tail-to-Tail interactions (Figs 2a, b and 3). This explains why they grow bi-directionally²¹. Moreover, the fact that both ends of NM are involved in prion propagation explains why rare cross-seeding events (between NM proteins from diverse species and between entirely different prions in the same species) rapidly convert from heterotypic to homotypic interactions^{35–37}. Once the heterologous protein has joined the fibre, it will immediately present a homotypic interface for either Head-to-Head or Tail-to-Tail interactions.

Our work also sheds light on the role played by the 5.5 degenerate oligopeptide repeats (consensus sequence P/QQGGYQQ/SYN). Increasing the number of repeats greatly increases spontaneous prion formation *in vivo* and fibre nucleation *in vitro*³⁸. Reducing them eliminates the protein's ability to propagate *in vivo* as a prion on its own. However, even a single repeat allows efficient incorporation into pre-existing wild-type prions^{38–40}. We find that most repeats are in the Central Core, sequestered from intermolecular contacts (Figs 2 and 4d), and that the length of the Central Core determines fibre stability (Fig. 5a). However, intermolecular Tail-to-Tail contacts include the final 1.5 to 2.5 repeats. Repeat-to-repeat contacts are therefore accommodated in both intramolecular and intermolecular amyloid interactions. With a single repeat, the stability of NM amyloid might fall below the threshold required to maintain the

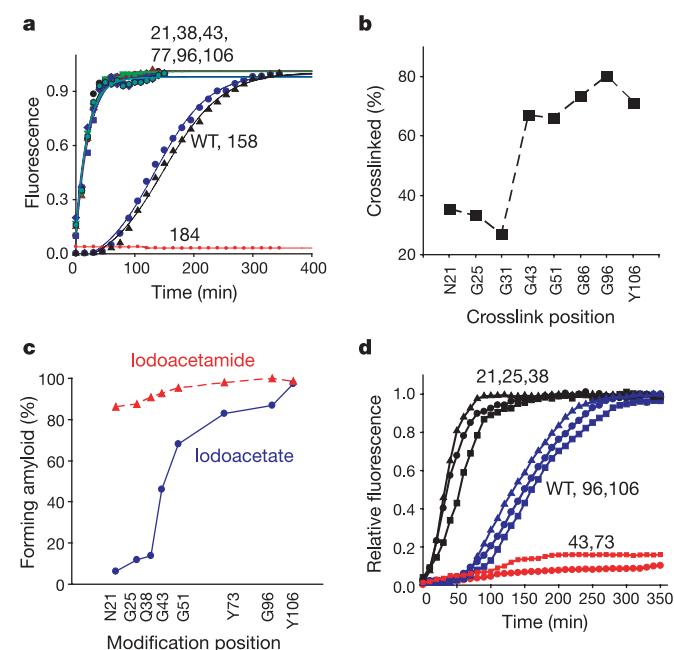


Figure 4 | Early assembly events. **a**, Formation of a collapsed intermediate. An increase in acrylodan fluorescence at positions N21 (black squares), Q38 (blue squares), G43 (red triangles), Y77 (green triangles), G96 (green hexagons) and Y106 (blue diamonds) occurred very rapidly after proteins were diluted into buffer, before the formation of amyloid (see also Supplementary Fig. S4). Amyloid formation was assessed by SDS insolubility because acrylodan and ThT fluorescence could not be analysed in the same reaction. Acrylodan fluorescence for mutants T158 (black triangles) and K184 (red circles) in the M domain, and ThT binding profiles of wild-type NM (blue circles) are plotted for comparison. **b**, Susceptibility of different cysteine residues to spontaneous crosslinking during assembly under non-reducing conditions. Crosslinks were assessed by non-reducing SDS–PAGE. **c**, Addition of a single charge in the Head region severely impedes fibre assembly. Assembly of negatively charged iodoacetate-labelled (blue) and uncharged iodoacetamide-labelled (red) proteins was monitored by ThT fluorescence. **d**, Assembly kinetics of cysteine variants crosslinked with BMB at position N21 (black triangles), G25 (black circles), Q38 (black squares), wild-type NM (blue triangles), G96 (blue circles), Y106 (blue squares), G43 (red squares) and Y73 (red circles), monitored by ThT fluorescence.

prion state on its own (perhaps through increased susceptibility to Hsp104 disaggregation²³). However, with this single repeat, the molecule could still join pre-existing wild-type prions through Head-to-Head and repeat-to-repeat interactions. Noting that the repeat sequence is able to form an amyloid on its own⁴¹, we further propose that the closely spaced repeat sequences of individual NM molecules readily collapse into transient pre-amyloid structures within the molten intermediates where nucleation occurs. The more repeats, the more likely this transient structure is to form, reducing the number of contacts the Head region must sample to initiate nucleation, and thereby increasing the spontaneous rate of prion formation.

What structural elements are responsible for enciphering prion strains? Previous data are perplexing. On one hand, amyloids assembled *in vitro* from a fragment containing only the first 65 residues of N can induce a broad spectrum of distinct heritable prion strains when used to transform [*psi*⁻] cells to [*PSI*⁺]²⁶. On the other hand, during *in vivo* propagation, residues 1–137 are required to maintain several [*PSI*⁺] strains⁴². The extremely reproducible pyrene excimer signals in our experiments (Figs 2a and 5b) provide a key. Different intermolecular contacts in both the Head and Tail create distinct strains. Residues 1–65 contain the Head region and two repeats. This fragment should be able to form the full spectrum of Head region interactions that distinguish different strains, and be able to propagate these to wild-type protein. However, differences in the length of the Central Core and the position of Tail-to-Tail

contacts are the major distinctions between prion strains. Hence, residues 65–137 are required for the maintenance of distinct strains *in vivo*.

Does our data have relevance to other prions and amyloids? Although it is not yet clear that molten oligomeric species are obligate intermediates in the formation of other amyloids (as it is for NM), they are present in the initial stages of many amyloid assemblies^{43,44}. Evidence also attests to the importance of dimeric interactions in the assembly of diverse amyloids^{45–47}. Recent molecular dynamics simulations with A β peptide suggest that assembly begins with a molten oligomeric species that allows the rapid sampling of multiple intermolecular contacts, with two individual molecules precipitously finding the right contact and nucleating assembly⁴⁴.

Several other similarities with the mammalian prion determinant, PrP, are apparent. These include (1) electron crystallography data supporting a β -helix-like structure in the prion state^{48,49}; (2) the importance of PHGGGWGQ oligopeptide repeats (an increase in repeat number causes spontaneous prion formation, and eliminating repeats inhibits propagation); and (3) the ability of PrP to form a variety of phenotypically distinct strains with different structures (that is, different sensitivities to GdmCl denaturation, protease cleavage sites and glycosylation states)⁵⁰.

For simplicity, we distinguished only strong and weak strains of NM but, in fact, each category includes a multiplicity of strains (ref. 27 and data not shown). We suggest that promiscuity in

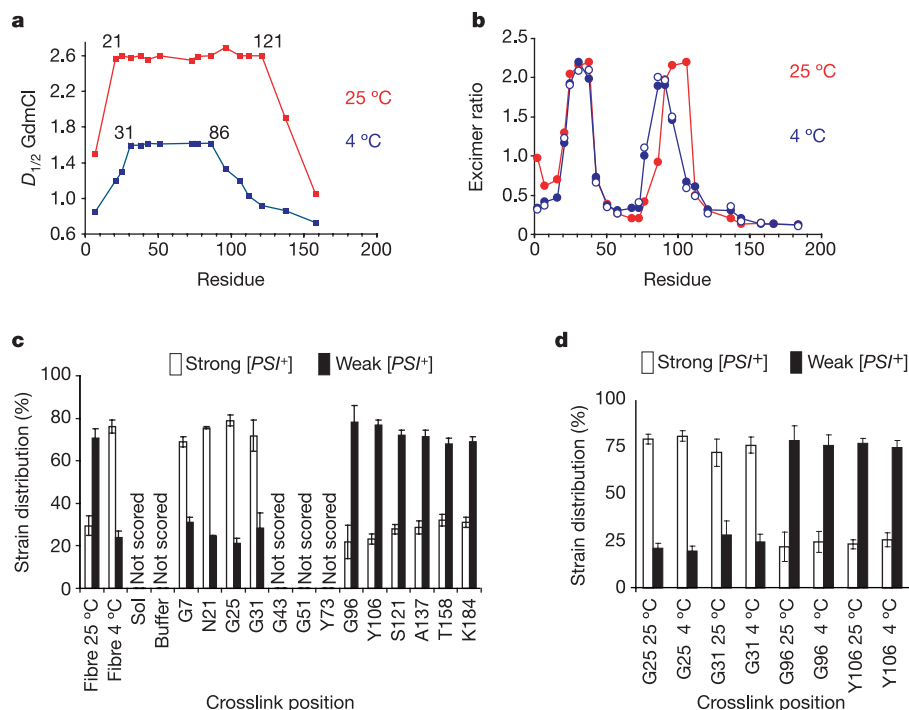


Figure 5 | Structural distinctions between prion strain populations.

a, Amyloid core length and stability differ in fibres assembled at 4°C and 25°C. Acrylodan-labelled fibres assembled at 4°C (blue) and 25°C (red) were denatured as in Fig. 1c. $D_{1/2}$ values were obtained from full GdmCl denaturation profiles (see Fig. 1c). **b**, Intersubunit interfaces change in fibres assembled at different temperatures. Excimer fluorescence of pyrene-labelled mutants assembled in rotated, unseeded reactions at 25°C (red), 4°C (blue) or in seeded, unrotated reactions with 4% seed formed at 4°C (open circles). Pyrene fluorescence is similar in seeded and unseeded reactions. **c**, BMB crosslinks at different positions bias assembly towards different prion strains. NM proteins crosslinked in denaturant were diluted into buffer, assembled into fibres, and used to transform [*psi*⁻] cells to the

[*PSI*⁺] state. Left columns show controls: un-crosslinked wild-type fibres assembled at 25°C or at 4°C are biased towards the production of weak or strong strains, respectively²⁷. Soluble NM, buffer alone, and proteins crosslinked in the Central Core did not induce [*PSI*⁺] above background and were not scored. Values represent means $\pm$ s.d. ($n = 5$ experiments). **d**, Strain biases created by crosslinking overcome the biases created by assembly at different temperatures. Whether they were assembled at 4°C or at 25°C, fibres cross-linked in the Head region (G25 or G31) produced mostly strong [*PSI*⁺] strains, and fibres crosslinked in the Tail region (G96 and Y106) produced mostly weak [*PSI*⁺] strains. Values represent means $\pm$ s.d. ($n = 5$ experiments).

intermolecular Head and Tail contacts, together with the low sequence complexity of the N domain (72% Q, N, G and Y) allow differences in intermolecular contacts to ratchet the amyloid core into different registers, creating an ensemble of structures with closely related but distinct β -strand contacts. If the length of the amyloid core and the residues involved in contact interfaces differ for PrP strains as they do for NM, it would readily explain variations in transmission within and between species. One fold might display an interface that is polymorphic between species, another might display an invariant interface. One fold might restrict the incorporation of a glycosylated residue because the modified side chain would point inward; another might permit it because the residue points outward. Similarly, transient intramolecular β -structures formed by expanded oligopeptide repeats would reduce the number of intermolecular contacts that must be sampled to nucleate prion formation, even though the prion-enhancing repeats of PrP lie outside the final amyloid core^{4,5}. There will undoubtedly be many variations in the structure and dynamics of amyloid assemblies, but several of the principles we have uncovered, as well as the methods we used, will probably be applicable to other systems.

Note added in proof: A study focusing on three residues (amino acids 36, 76 and 108) has recently reported that paramagnetic spin resonance spectroscopy provides evidence for strain-specific differences that affect the species barrier between *S. cerevisiae* and *C. albicans*⁵¹.

METHODS

Mutagenesis, protein purification and yeast strains. The site-specific substitution of individual amino-acid codons for cysteine codons, the production of integrating genomic constructs, the strategy for replacing wild-type SUP35 with the cysteine mutants, and overexpression and purification of cysteine mutants is described in the Supplementary Information.

Cysteine labelling. Post-assembly cysteine labelling: 5 μ M NM fibres were incubated with 15 μ M of pyrene maleimide or 15 μ M Lucifer yellow for 3 h. Fibres were washed three times with 20% methanol containing 5 mM DTT, and redissolved in 6 M GdmCl. Labelling efficiencies were calculated according to the Molecular probes website (www.probes.com).

Pre-assembly cysteine labelling: proteins were incubated in 6 M GdmCl with acrylodan or pyrene maleimide as recommended by Molecular Probes (see Supplementary Information). Labelling efficiencies for acrylodan (>90%) and pyrene (70–80%) were also determined according to the manufacturer's protocol.

Iodoacetate and iodoacetamide labelling in 6 M GdmCl used 25 μ M protein and a fivefold excess of iodoacetate or iodoacetamide for 2 h at 25 °C. After using a PD10 desalting column (Pharmacia) to remove free label, efficiencies were determined from the number of free thiol groups still accessible to 5,5'-dithiobis-2-nitrobenzoic acid.

BMB (1,4-bis-maleimidobutane) was used at a protein-to-reagent ratio of 1:2. O-DTT (oxidized dithiothreitol) was used at a ratio of 5:1. Reactions in 6 M GdmCl were terminated after 3 h with 5 mM DTT. Crosslinking efficiencies (all 70–85%) were assessed by SDS-PAGE.

Protein transformation. Yeast cells containing an ADE1 mutation suppressible by [PSI⁺] were co-transformed with NM proteins and a URA3 plasmid as described²⁷. Ura⁺ transformants were picked before colour development (three days instead of seven), to avoid bias. Approximately 200 transformants for each mutant (obtained from five independent transformations in each case), were patched onto uracil-deficient medium and replica plated to both adenine-deficient medium and 25% rich medium (YPD; yeast extract, peptone and dextrose) supplemented with 20 μ g ml⁻¹ adenine. Transformation efficiencies were calculated (Supplementary Fig. S6) and strong and weak strains were scored²⁷ by colour on YPD (Supplementary Fig. S1), and rates of growth were scored on adenine-deficient medium.

Received 30 December 2004; accepted 26 April 2005.

1. Tuite, M. F. & Cox, B. S. Propagation of yeast prions. *Nature Rev. Mol. Cell Biol.* **4**, 878–890 (2003).
2. Chien, P., Weissman, J. S. & DePace, A. H. Emerging principles of conformation-based prion inheritance. *Annu. Rev. Biochem.* **73**, 617–656 (2004).
3. Shorter, J. & Lindquist, S. Prions as adaptive conduits of memory and inheritance. *Nature Rev. Genet.* doi:10.1038/nrg1616 (in the press).

4. Caughey, B. Transmissible spongiform encephalopathies, amyloidoses and yeast prions: common threads? *Nature Med.* **6**, 751–754 (2000).
5. Prusiner, S. B. Prions. *Proc. Natl Acad. Sci. USA* **95**, 13363–13383 (1998).
6. Uptain, S. M. & Lindquist, S. Prions as protein-based genetic elements. *Annu. Rev. Microbiol.* **56**, 703–741 (2002).
7. Si, K. et al. A neuronal isoform of CPEB regulates local protein synthesis and stabilizes synapse-specific long-term facilitation in *Aplysia*. *Cell* **115**, 893–904 (2003).
8. Si, K., Lindquist, S. & Kandel, E. R. A neuronal isoform of the *Aplysia* CPEB has prion-like properties. *Cell* **115**, 879–891 (2003).
9. Dobson, C. M. Protein folding and misfolding. *Nature* **426**, 884–890 (2003).
10. Jensen, M. A., True, H. L., Chernoff, Y. O. & Lindquist, S. Molecular population genetics and evolution of a prion-like protein in *Saccharomyces cerevisiae*. *Genetics* **159**, 527–535 (2001).
11. Eaglestone, S. S., Cox, B. S. & Tuite, M. F. Translation termination efficiency can be regulated in *Saccharomyces cerevisiae* by environmental stress through a prion-mediated mechanism. *EMBO J.* **18**, 1974–1981 (1999).
12. True, H. L. & Lindquist, S. L. A yeast prion provides a mechanism for genetic variation and phenotypic diversity. *Nature* **407**, 477–483 (2000).
13. True, H. L., Berlin, I. & Lindquist, S. L. Epigenetic regulation of translation reveals hidden genetic variation to produce complex traits. *Nature* **431**, 184–187 (2004).
14. Kushnirov, V. V. et al. Divergence and conservation of SUP2 (SUP35) gene of yeast *Pichia pinus* and *Saccharomyces cerevisiae*. *Yeast* **6**, 461–472 (1990).
15. Ter-Avanesyan, M. D. et al. Deletion analysis of the SUP35 gene of the yeast *Saccharomyces cerevisiae* reveals two non-overlapping functional regions in the encoded protein. *Mol. Microbiol.* **7**, 683–692 (1993).
16. Chernoff, Y. O., Derkach, I. L. & Inge-Vechtomov, S. G. Multicopy SUP35 gene induces de-novo appearance of psi-like factors in the yeast *Saccharomyces cerevisiae*. *Curr. Genet.* **24**, 268–270 (1993).
17. Glover, J. R. et al. Self-seeded fibers formed by Sup35, the protein determinant of [PSI⁺], a heritable prion-like factor of *S. cerevisiae*. *Cell* **89**, 811–819 (1997).
18. Liu, J. J., Sondheimer, N. & Lindquist, S. L. Changes in the middle region of Sup35 profoundly alter the nature of epigenetic inheritance for the yeast prion [PSI⁺]. *Proc. Natl Acad. Sci. USA* **99** (suppl. 4), 16446–16453 (2002).
19. Li, L. & Lindquist, S. Creating a protein-based element of inheritance. *Science* **287**, 661–664 (2000).
20. Serio, T. R. et al. Nucleated conformational conversion and the replication of conformational information by a prion determinant. *Science* **289**, 1317–1321 (2000).
21. Scheibel, T., Kowal, A. S., Bloom, J. D. & Lindquist, S. L. Bidirectional amyloid fiber growth for a yeast prion determinant. *Curr. Biol.* **11**, 366–369 (2001).
22. Scheibel, T., Bloom, J. & Lindquist, S. L. The elongation of yeast prion fibers involves separable steps of association and conversion. *Proc. Natl Acad. Sci. USA* **101**, 2287–2292 (2004).
23. Shorter, J. & Lindquist, S. Hsp104 catalyzes formation and elimination of self-replicating Sup35 prion conformers. *Science* **304**, 1793–1797 (2004).
24. Collins, S. R., Douglass, A., Vale, R. D. & Weissman, J. S. Mechanism of prion propagation: amyloid growth occurs by monomer addition. *PLoS Biol.* **2**, e321 (2004).
25. DePace, A. H. & Weissman, J. S. Origins and kinetic consequences of diversity in Sup35 yeast prion fibers. *Nature Struct. Biol.* **9**, 389–396 (2002).
26. King, C. Y. & Diaz-Avalos, R. Protein-only transmission of three yeast prion strains. *Nature* **428**, 319–323 (2004).
27. Tanaka, M., Chien, P., Naber, N., Cooke, R. & Weissman, J. S. Conformational variations in an infectious protein determine prion strain differences. *Nature* **428**, 323–328 (2004).
28. Kishimoto, A. et al. β -Helix is a likely core structure of yeast prion Sup35 amyloid fibers. *Biochem. Biophys. Res. Commun.* **315**, 739–745 (2004).
29. DePace, A. H., Santos, A., Hillner, P. & Weissman, J. S. A critical role for amino-terminal glutamine/asparagine repeats in the formation and propagation of a yeast prion. *Cell* **93**, 1241–1252 (1998).
30. King, C. Y. Supporting the structural basis of prion strains: induction and identification of [PSI] variants. *J. Mol. Biol.* **307**, 1247–1260 (2001).
31. Kajava, A. V., Baxa, U., Wickner, R. B. & Steven, A. C. A model for Ure2p prion filaments and other amyloids: the parallel superpleated β -structure. *Proc. Natl Acad. Sci. USA* **101**, 7885–7890 (2004).
32. Kamen, D. E., Griko, Y. & Woody, R. W. The stability, structural organization, and denaturation of pectate lyase C, a parallel β -helix protein. *Biochemistry* **39**, 15932–15943 (2000).
33. Santos, A., Chien, P., Osherovich, L. Z. & Weissman, J. S. Molecular basis of a yeast prion species barrier. *Cell* **100**, 277–288 (2000).
34. Chien, P., DePace, A. H., Collins, S. R. & Weissman, J. S. Generation of prion transmission barriers by mutational control of amyloid conformations. *Nature* **424**, 948–951 (2003).
35. Bradley, M. E., Edskes, H. K., Hong, J. Y., Wickner, R. B. & Liebman, S. W. Interactions among prions and prion “strains” in yeast. *Proc. Natl Acad. Sci. USA* **99** (suppl. 4), 16392–16399 (2002).
36. Derkach, I. L. et al. Effects of Q/N-rich, polyQ, and non-polyQ amyloids on the de novo formation of the [PSI⁺] prion in yeast and aggregation of Sup35 in vitro. *Proc. Natl Acad. Sci. USA* **101**, 12934–12939 (2004).

37. Nakayashiki, T., Ebihara, K., Bannai, H. & Nakamura, Y. Yeast [PSI⁺] “prions” that are crosstransmissible and susceptible beyond a species barrier through a quasi-prion state. *Mol. Cell* **7**, 1121–1130 (2001).
38. Liu, J. J. & Lindquist, S. Oligopeptide-repeat expansions modulate ‘protein-only’ inheritance in yeast. *Nature* **400**, 573–576 (1999).
39. Parham, S. N., Resende, C. G. & Tuite, M. F. Oligopeptide repeats in the yeast protein Sup35p stabilize intermolecular prion interactions. *EMBO J.* **20**, 2111–2119 (2001).
40. Osherovich, L. Z., Cox, B. S., Tuite, M. F. & Weissman, J. S. Dissection and design of yeast prions. *PLoS Biol.* **2**, E86 (2004).
41. Balbirnie, M., Grothe, R. & Eisenberg, D. S. An amyloid-forming peptide from the yeast prion Sup35 reveals a dehydrated β -sheet structure for amyloid. *Proc. Natl Acad. Sci. USA* **98**, 2375–2380 (2001).
42. Bradley, M. E. & Liebman, S. W. The Sup35 domains required for maintenance of weak, strong or undifferentiated yeast [PSI⁺] prions. *Mol. Microbiol.* **51**, 1649–1659 (2004).
43. Eakin, C. M., Attenello, F. J., Morgan, C. J. & Miranker, A. D. Oligomeric assembly of native-like precursors precedes amyloid formation by β -2 microglobulin. *Biochemistry* **43**, 7808–7815 (2004).
44. Hwang, W., Zhang, S., Kamm, R. D. & Karplus, M. Kinetic control of dimer structure formation in amyloid fibrillogenesis. *Proc. Natl Acad. Sci. USA* **101**, 12916–12921 (2004).
45. Lee, S. & Eisenberg, D. Seeded conversion of recombinant prion protein to a disulfide-bonded oligomer by a reduction-oxidation process. *Nature Struct. Biol.* **10**, 725–730 (2003).
46. Garzon-Rodriguez, W. *et al.* A conformation change in the carboxyl terminus of Alzheimer’s A β (1–40) accompanies the transition from dimer to fibril as revealed by fluorescence quenching analysis. *J. Biol. Chem.* **275**, 22645–22649 (2000).
47. Barghorn, S. & Mandelkow, E. Toward a unified scheme for the aggregation of tau into Alzheimer paired helical filaments. *Biochemistry* **41**, 14885–14896 (2002).
48. Wille, H. *et al.* Structural studies of the scrapie prion protein by electron crystallography. *Proc. Natl Acad. Sci. USA* **99**, 3563–3568 (2002).
49. Govaerts, C., Wille, H., Prusiner, S. B. & Cohen, F. E. Evidence for assembly of prions with left-handed β -helices into trimers. *Proc. Natl Acad. Sci. USA* **101**, 8342–8347 (2004).
50. Safar, J. *et al.* Eight prion strains have PrP^{Sc} molecules with different conformations. *Nature Med.* **4**, 1157–1165 (1998).
51. Tanaka, M., Chien, P., Yonekura, K. & Weissman, J. S. Mechanism of cross-species prion transmission: an infectious conformation compatible with two highly divergent yeast prion proteins. *Cell* **121**, 49–62 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Weissman for providing us with the fibre transformation protocols, members of the Lindquist, Langen and Berger laboratories for helpful discussions, and R. Latek and T. DiCesare for help with the figures. This research was supported by the DuPont-MIT Alliance and an NIH grant.

Author Contributions Experimental work was performed by R.K., data analysis by R.K. and S.L.L., and writing primarily by S.L.L.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.L.L. (Lindquist_admin@wi.mit.edu).

Structure of the cross- β spine of amyloid-like fibrils

Rebecca Nelson¹, Michael R. Sawaya¹, Melinda Balbirnie¹, Anders Ø. Madsen^{2,3}, Christian Riek³, Robert Grothe¹ & David Eisenberg¹

Numerous soluble proteins convert to insoluble amyloid-like fibrils that have common properties. Amyloid fibrils are associated with fatal diseases such as Alzheimer's, and amyloid-like fibrils can be formed *in vitro*. For the yeast protein Sup35, conversion to amyloid-like fibrils is associated with a transmissible infection akin to that caused by mammalian prions. A seven-residue peptide segment from Sup35 forms amyloid-like fibrils and closely related microcrystals, from which we have determined the atomic structure of the cross- β spine. It is a double β -sheet, with each sheet formed from parallel segments stacked in register. Side chains protruding from the two sheets form a dry, tightly self-complementing steric zipper, bonding the sheets. Within each sheet, every segment is bound to its two neighbouring segments through stacks of both backbone and side-chain hydrogen bonds. The structure illuminates the stability of amyloid fibrils, their self-seeding characteristic and their tendency to form polymorphic structures.

Four decades of research have established that amyloid-like fibrils of different proteins have a common structural 'cross- β ' spine¹. In 1959, elongated, unbranched fibrils were observed in electron micrographs of diseased tissues², and in 1968 came the discovery that fibrils exhibit an X-ray diffraction signature known as the cross- β pattern³. This pattern shows⁴ that the strongest repeating feature of the fibril is a set of β -sheets that are parallel to the fibril axis, with their strands perpendicular to this axis. The hypothesis of a common molecular organization was supported by the finding⁵ that amyloid fibrils from six different proteins, each associated with its own clinical syndrome, showed similar cross- β diffraction patterns. The degree of similarity pointed to a common core molecular structure.

Revealing the atomic details of this cross- β spine has been impeded by the limited order of fibrils isolated from diseased tissues, infected cells and *in vitro* conversions of proteins to fibrils. There is also evidence for a diversity of crystalline and fibril structures^{6–8}. Nevertheless, an array of biophysical tools has defined important features. These tools include solid-state NMR^{9–11}, model-building constrained by X-ray fibre and powder diffraction^{6,7,12,13}, site-directed spin labelling^{14,15}, cryo-electron microscopy^{16,17}, and proline-scanning mutagenesis¹⁸. Despite numerous models suggested by these studies, until now no refined, fully objective atomic model has been available for the common spine structure.

We selected the yeast protein Sup35 for X-ray diffraction analysis because extensive studies have shown that its fibril formation is the basis of protein-based inheritance and prion-like infectivity^{19–23}. Its fibril-forming tendency had been traced to the amino terminus of the prion-determining domain^{24,25}, and from this region we isolated a seven-residue, fibril-forming segment with sequence GNNQQNY⁶. This peptide dissolves in water, and at a concentration of approximately 400 μ M, forms amyloid-like fibrils in a few hours. These fibrils display all of the common characteristics of amyloid fibrils, including: elongated, unbranched morphology; the cross- β diffraction pattern; binding of the flat dyes Congo red and thioflavin T; the characteristic green–yellow birefringence of Congo red;

lag-dependent cooperative kinetics of formation with self-seeding²⁶; and unusual stability.

GNNQQNY and the related peptide NNQQNY form elongated microcrystals at higher concentrations (about 10–100 mM), enabling X-ray diffraction studies. The microcrystals are similar to the fibrils in that the peptide segments are perpendicular to the long dimension of both aggregates and that fibrils and microcrystals have similar diffraction patterns (Supplementary Fig. 2). In hundreds of crystallization experiments, microcrystals never grew to more than a few micrometres in length, with much narrower cross-sections.

Architecture of the GNNQQNY cross- β spine

Three features of the microcrystals made it possible to determine structures for GNNQQNY and NNQQNY. First, the largest microcrystals (Fig. 1) are of sufficient size, order and stability to yield adequate diffraction data on microfocus beamline ID13 at the European Synchrotron Radiation Facility (ESRF). Second, microcrystals of NNQQNY grow only in the presence of Zn²⁺ or Cd²⁺. Anomalous scattering from a crystal of Zn-NNQQNY yielded phases for the structure of Zn-NNQQNY. Third, the structure of GNNQQNY is nearly isomorphous with that of NNQQNY, allowing structure determination from a difference map. Details of data collection and structure determination are listed in Table 1. The NNQQNY structure is described in Supplementary Information. Here we focus on the structure of GNNQQNY.

GNNQQNY molecules are extended in conformation and are hydrogen-bonded to each other in standard Pauling–Corey parallel β -sheets (Supplementary Table 1). Because the strands are perpendicular to the long axis of the microcrystals, hydrogen-bonded addition of GNNQQNY molecules to the growing β -sheet accounts for the elongated shape of the crystals as well as the fibrils. As previously suggested from X-ray powder diffraction of the microcrystals^{6,7}, the GNNQQNY β -strands within each sheet are parallel and exactly in register (Fig. 2a). A parallel, in register arrangement is also seen for A β molecules in their fibrils^{9,10}. Each pair of sheets is

¹Howard Hughes Medical Institute, UCLA-DOE Institute for Genomics and Proteomics, Box 951570, UCLA, Los Angeles, California 90095-1570, USA. ²Centre for Crystallographic Studies, Department of Chemistry, University of Copenhagen, Universitetsparken 5, DK-2100 KBH, Denmark. ³European Synchrotron Radiation Facility, B.P. 220, F-38043 Grenoble Cedex, France.

related by a 2_1 screw axis: the strands in one sheet are antiparallel to those in the mating sheet, and each sheet is shifted along the screw axis relative to its mate by one-half the strand-strand separation of 4.87 Å. Thus, side chains extending from a strand in one sheet nestle between side chains extending from two strands of the mating sheet (Fig. 2b).

There are two distinctly different interfaces between sheets, which we term the dry and wet interfaces (Fig. 2c). The wet interface is lined with water molecules that completely separate GNNQQNY molecules, other than a contact between Tyr 7 residues in neighbouring sheets. The separation of these sheets is large, about 15 Å. In contrast, the dry interface contains no water, other than two molecules that hydrate the carboxylate ions at the ends of the peptide segments. These sheets are closer together, separated by 8.5 Å. Whereas each polar side chain of the wet interface is hydrated by water molecules, the polar side chains of the dry interface (Asn 2, Gln 4, and Asn 6) are tightly interdigitated with the same three side chains of the mating sheet (Fig. 2d). These opposing side chains do not form hydrogen bonds with each other; rather, their shapes complement each other closely, forming van der Waals interactions. Viewed down the sheets (Fig. 2d) the interdigitating side chains look like the teeth of a zipper, so we call this interaction a 'steric zipper'. The dry interface is a stack of these steric zippers (Fig. 2b).

The shape complementarity of the dry interface is unusually tight when compared to other protein interfaces, as quantified by the S_C parameter²⁷. S_C measures the shape complementarity of two atomic surfaces by comparing the directions of unit vectors normal to the two surfaces, emanating from nearest points on the opposed surfaces. The average dot product of the pair of vectors approaches 1.0 as the two surfaces follow each other perfectly. The tightly meshing surfaces of proteolytic inhibitor proteins and their cognate proteases have values of S_C in the range 0.73 ± 0.03 , and S_C for protein antigens bound to antibodies are 0.66 ± 0.02 (ref. 27). For the sheets forming the steric zipper in the dry interface $S_C = 0.86$, showing that this interface has unusually high complementarity.

The remarkable complementarity between sheets in the dry interface suggests that the stable structural unit of the cross- β spine is a pair of β -sheets. The wet interface, with only a single peptide-peptide contact, has the features of a crystal contact and may not exist in the fibril structure. A pair-of-sheets organization for the cross- β spine is consistent with several other observations. First, a spine of two sheets is self-limiting in lateral growth, because the same face of both sheets is opposed, exposing a different outward face—in this structure, the wet face. A spine of three or four such stacked sheets would expose a

face identical to one of its interior bonding faces, leading to further lateral growth. Second, models of cross- β spines containing three or more sheets sustain distortions in backbone hydrogen bonding that increase as the sheets stack farther from the fibril axis. Third, the width of the diffuse equatorial X-ray reflection at approximately 9–11 Å resolution in fibrils of β 2-microglobulin corresponds better with a model containing two sheets than a model containing a single sheet or three sheets²⁸. Finally, a pair-of-sheets structure is consistent with studies by cryo-electron microscopy of the amyloid-like protofibrils of SH3 and insulin^{16,29}. In short, the crystal structures of GNNQQNY and NNQQNY suggest that a tight, dry steric fit between a pair of sheets is likely to be a fundamental feature of amyloid-like fibrils. However, it is not yet clear how to reconcile a pair-of-sheets feature with evidence from mass-per-unit-length measurements on A β fibrils¹⁰ and from electron microscopy measurements of GNNQQNY protofibrils⁷, which are consistent with four sheets.

Another fundamental feature of the cross- β spine shown in Fig. 2 is that it is built from a short peptide. The self-complementary steric zipper explains how short segments of proteins are able to form amyloid-like fibrils and raises the question of whether the rest of the protein participates in the spine.

Amide stacks in the cross- β spine

Although there are no hydrogen bonds bridging two tightly complementing sheets across the dry interface, each GNNQQNY molecule forms 11 hydrogen bonds to its two neighbouring molecules in the same sheet (Fig. 2e). Five of these are backbone C=O-H-N hydrogen bonds, and four are 'amide stacks'; that is, amide-amide hydrogen bonds between pairs of identical Asn or Gln residues in adjacent molecules within a sheet. It is these hydrogen-bonded amide stacks that force the GNNQQNY and NNQQNY molecules to stack parallel and in register in their respective sheets. This network of backbone and side-chain hydrogen bonds is reminiscent of the polar zipper proposed previously³⁰. Amide stacks such as those found here could stabilize the polyglutamine aggregates formed in the CAG expansion diseases and those formed *in vitro* with polyglutamine-containing peptides. The remaining hydrogen bonds between

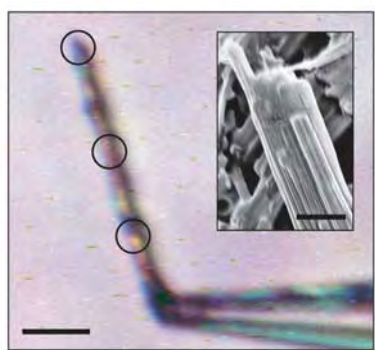


Figure 1 | The NNQQNY microcrystal used for X-ray diffraction data collection, held to the tip of a glass capillary by cryoprotectant (50% ethylene glycol/water). Scale bar, 10 μ m. X-rays were focused on the encircled areas. Separate data sets were collected for each and were merged to provide the final data set. The inset shows a scanning electron micrograph of NNQQNY crystals, suggesting that the 'large' microcrystals used for data collection are composed of several aligned, nanometre-sized blocks. Scale bar of inset, 1 μ m.

Table 1 | Statistics of data collection, phasing and atomic refinement

Data collection	NNQQNY*	GNNQQNY
Space group	$P2_1$	$P2_1$
Resolution (Å)	1.3	1.8
Unit cell dimensions		
<i>a</i> (Å)	21.15	21.94
<i>b</i> (Å)	4.87	4.87
<i>c</i> (Å)	23.13	23.48
β (°)	102.93	107.08
Measured reflections	8,241	995
Unique reflections	2,166	509
Overall completeness (%)	97.1	89.5
Last shell completeness (%)	88.8	84.2
Overall R_{sym} †	0.146	0.204
Last shell R_{sym}	0.426	0.491
Overall $I/\sigma(I)$	9.9	3.8
Last shell $I/\sigma(I)$	2.6	1.5
Refinement		
R_{work}	0.102	0.181
R_{free}	0.152	0.190
r.m.s.d. bond length (Å)	0.007	0.014
r.m.s.d. bond angle (°)	1.1	1.2
Number of protein atoms	55	59
Number of solvent atoms	12	7
Average <i>B</i> factor of protein atoms	5.6	13.1
Average <i>B</i> factor of solvent atoms	17.6	27.5
PDB ID code	1yjo	1yjp

r.m.s.d., root mean square deviation.

*Friedel pairs are not merged in reported NNQQNY data statistics. The NNQQNY structure was refined with anisotropic *B* factors.

† $R_{\text{sym}}(I) = \sum_h \sum_l (|I_{hkl}| - \langle I_{hkl} \rangle) / \sum_l |I_{hkl}|$.

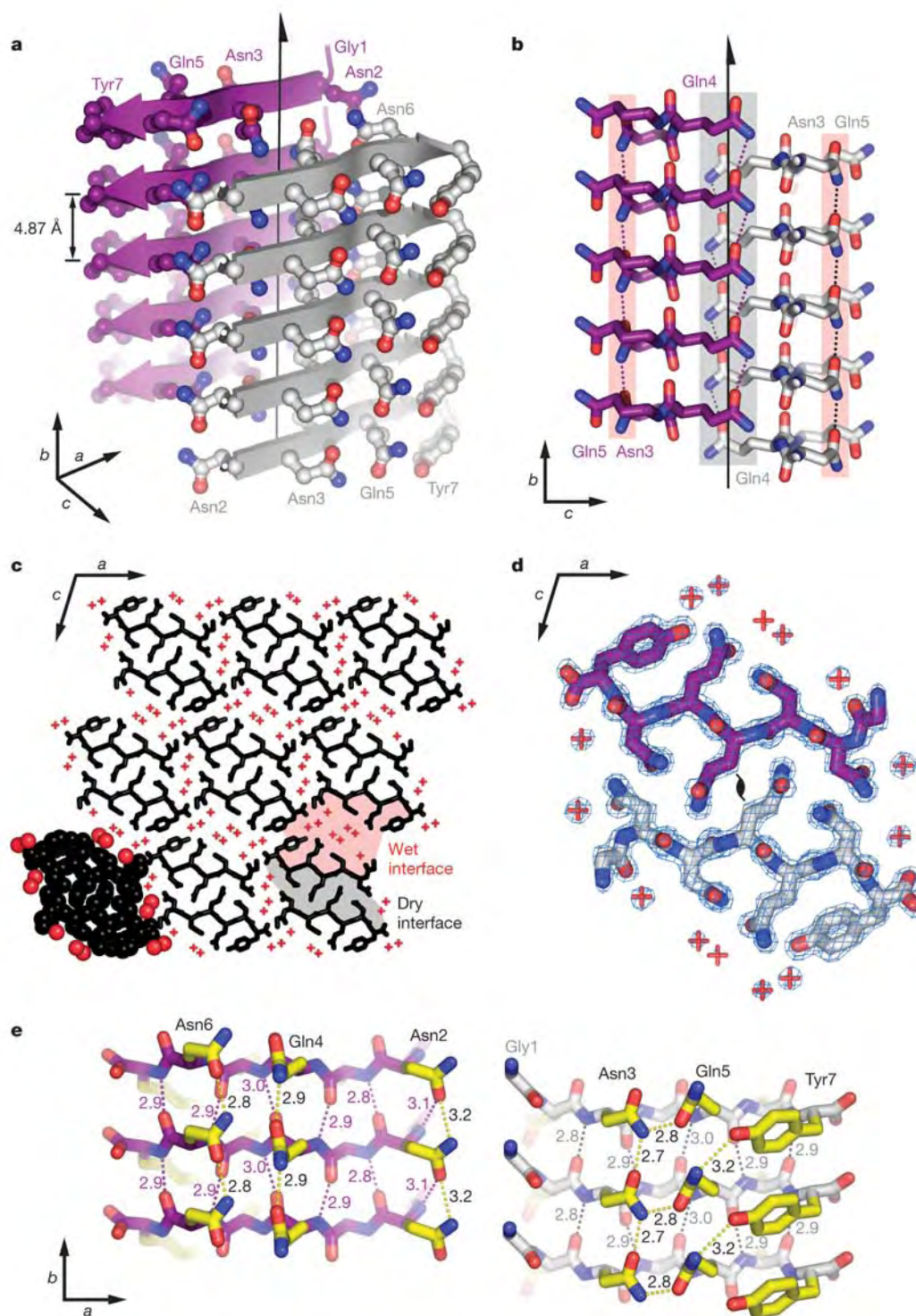


Figure 2 | Structure of GNNQQNY. Unless otherwise noted, carbon atoms are coloured in purple or grey/white, oxygen in red, nitrogen in blue. **a**, The pair-of-sheets structure, showing the backbone of each β -strand as an arrow, with side chains protruding. The dry interface is between the two sheets, with the wet interfaces on the outside surfaces. Side chains Asn 2, Gln 4 and Asn 6 point inwards, forming the dry interface. The 2_1 screw axis of the crystal is shown as the vertical line. It rotates one of the strands of the near sheet 180° about the axis and moves it up $4.87 \text{ \AA}/2$ so that it is superimposed on one of the strands of the far sheet. **b**, The steric zipper viewed edge on (down the a axis). Note the vertical shift of one sheet relative to the other, allowing interdigitation of the side chains emanating from each sheet. The amide stacks of the dry interface are shaded in grey at the centre, and those of the wet interface are shaded in pale red on either side. **c**, The GNNQQNY crystal viewed down the sheets (from the top of panel **a**, along the b axis). Six

rows of β -sheets run horizontally. Peptide molecules are shown in black and water molecules are red plus signs. The atoms in the lower left unit cell are shown as spheres representing van der Waals radii. **d**, The steric zipper. This is a close-up view of a pair of GNNQQNY molecules from the same view as panel **c**, showing the remarkable shape complementarity of the Asn and Gln side chains protruding into the dry interface. $2F_o - F_c$ electron density is shown, and the position of the central screw axis is indicated. **e**, Views of the β -sheets from the side (down the c axis), showing three β -strands with the inter-strand hydrogen bonds. Side-chain carbon atoms are yellow. Backbone hydrogen bonds are shown by purple or grey dots and side-chain hydrogen bonds by yellow dots. Hydrogen bond lengths are noted in $\text{\AA}$. The views of the interfaces are close to the views of panel **a**. The left-hand set is viewed from the centre of the dry interface; the right-hand set is viewed from the wet interface. Note the amide stacks in both interfaces.

GNNQQNY molecules in the sheet are from the side-chain nitrogen of Gln 5 to the hydroxyl of Tyr 7 and from the Asn 2 backbone nitrogen to the Asn 2 side-chain oxygen. Also, the rings of Tyr 7 are stacked, but not face-to-face: they pack edge-to-face across the wet interface.

Similarity to other structures

The structure of the GNNQQNY cross- β spine shows limited similarity to β -helices proposed as models for amyloid and prion spines^{17,31–34}. A search for structurally related β -sandwiches in the Protein Data Bank (PDB) yielded only one significant match to the backbone of GNNQQNY: SufD (PDB entry 1vh4). The search model, which contains six strands of GNNQQNY, three from each sheet forming the dry interface, can be superimposed on the SufD backbone with a root mean square (r.m.s.) deviation of 1.8 Å. The fold of SufD, a member of the β -helix family, resembles GNNQQNY more closely than do canonical right-handed β -helices because it has two sheets rather than three, and its sheets abut rather than having a cylindrical or triangular cross-section. SufD's similarity to GNNQQNY is limited, however, by its lack of a steric zipper; side chains from opposing sheets contact but do not interdigitate. As a result, the distance between sheets in SufD is nearly 2 Å greater. Hence, the complementarity between the two sheets composing SufD ($S_C = 0.70$) is significantly lower than in GNNQQNY. In short, the GNNQQNY structure shows only weak similarity to β -helices in general, and differs considerably from the cylindrical and triangular β -helices that have been proposed as models for amyloid-like spines.

Structure-based energetics

The structure of GNNQQNY suggests factors that determine the rate and stability of fibril formation as well as a factor that may underlie amyloid fibril polymorphism and prion strains^{8,22,23}. The structure indicates three levels of organization within the fibrils. The first is the alignment of GNNQQNY molecules to form a β -sheet. The second is the self-complementation of two sheets, forming the pair-of-sheets structure, with a dry interface. Because the self-complementation of two sheets involves van der Waals forces rather than hydrogen bonding, the patterns of bonding are less specific than those of the first level. Alternative interdigitations could give rise to fibril poly-

morphism and prion strains. In the third level, pair-of-sheets structures interact to form a fibril. For the third level, we note only that the non-covalent forces involved are probably weaker than those driving the formation of the first two levels.

For the alignment of GNNQQNY molecules to form a β -sheet, each GNNQQNY molecule must be extended. Because β -sheets form rapidly^{35,36} and reversibly, we assume that this level forms more rapidly than the second level. The second level is likely to form more slowly because the amide side chains must acquire the proper rotamers to permit interdigitation with the mating sheet and must be dehydrated to permit formation of the dry amide-stacking hydrogen bonds. We suggest that the decrease in entropy accompanying this step creates the barrier to fibril formation, which is evident in its lag-dependent cooperative formation. Once a nucleus of the cross- β spine has formed, additional molecules can be added more readily, leading to rapid growth. In the Supplementary Information we use the structure to argue that the nucleus for GNNQQNY fibril formation is about four molecules, and that the transition-state complex on the path to the nucleus is approximately three molecules. From energetic considerations we estimate a crude value for the free energy of forming this complex of $\sim 8 \text{ kcal mol}^{-1}$ of GNNQQNY at room temperature. If there are three molecules in the transition-state complex, the barrier is $\sim 24 \text{ kcal mol}^{-1}$, a substantial barrier to fibril formation.

In the formation of the transition-state complex and of the protofibril itself, there must be enthalpy decreases that compensate for the entropy decreases. Some enthalpy will be released by the van der Waals energy of the tight interdigitation in the steric zipper. The formation of hydrogen bonds between backbone groups and amide stacks will also contribute, but these bonds replace hydrogen bonds between water and the peptide in solution, so there is little net increase in the number of hydrogen bonds³⁷. Conceivably, hydrogen bonds in the pair-of-sheets structure are stronger than those in solution. They are in an anhydrous, low dielectric constant environment, and the columns of hydrogen bonds in the amide stacks run antiparallel to neighbouring columns (Fig. 2e), so there could be substantial strengthening of hydrogen bonds through induced dipoles, as is the case in ice³⁸. Although our estimates are crude, the standard free energy change for protofibril formation, ΔG^0 , the sum of the enthalpic (ΔH^0) and entropic terms (ΔS^0), is unlikely to be strongly negative.

Amyloid-like fibrils are stabilized by protein concentration as well as by formation of the steric zipper and the hydrogen bond stacks. For conversion of n peptide monomers, M , to an amyloid spine, M_n , with infinite cooperativity, $nM \rightarrow M_n$, the free energy of transition from the dissolved to the aggregated state is given by

$$\Delta G = \Delta G^0 + RT \ln \frac{[M_n]}{[M]^n} = \Delta H^0 - T\Delta S^0 + RT \ln \frac{[M_n]}{[M]^n}$$

in which ΔG^0 is the standard free energy, RT is the product of the gas constant with the absolute temperature, and the term on the right is governed by the concentration of monomer. At high concentrations of monomer, this term is strongly negative, favouring transition to the fibrillar state. Thus, our structure suggests that there is a large entropic barrier to amyloid fibril formation, but once a nucleus is present, high concentrations of protein drive the formation and contribute to an even larger barrier to dissolution of the fibrils (Fig. 3).

Summary and biological implications

The structures of GNNQQNY and NNQQNY determined here by X-ray microcrystallography confirm gross features of the cross- β spine that have been known from other methods: the spine is built from β -strands that are spaced approximately 4.8 Å apart, perpendicular to the fibril axis, formed into β -sheets with hydrogen bonds parallel to the axis, and exactly in register^{6,9,10}. What is new is the pair-of-sheets organization, with the interface between the paired sheets

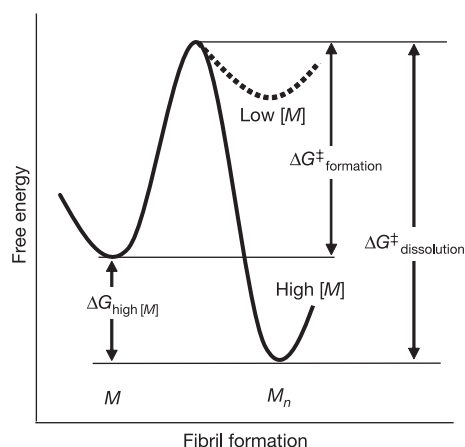


Figure 3 | A conjectural plot of the free energy, G , for conversion of monomeric GNNQQNY, M , to the aggregated state, M_n . The standard free energy change ΔG^0 for the conversion is small, so that the change in ΔG is controlled mainly by the concentration of monomer. At low concentrations, the monomeric state is favoured over the aggregated state, and the aggregated state is favoured at high concentrations. There is a significant kinetic barrier to formation of the aggregated state, $\Delta G_{\text{formation}}^\ddagger$. At high concentrations of protein, the barrier to re-dissolve fibres, $\Delta G_{\text{dissolution}}^\ddagger$, is very large.

consisting of the closely enmeshed self-complementing side chains protruding from the two sheets, termed a steric zipper. This interface is dry, in contrast to the highly hydrated external faces of the paired sheets. Disruption or capping of this steric zipper may be a strategy for drug interference of amyloid formation³⁹.

The steric zipper in the structures of GNNQQNY and NNQQNY explains how a fibril can be formed from a short segment of a protein. In fact, fibrils formed from short peptides are well known^{6,40,41}. We suggest that such short segments are capable of self-complementation across a dry inter-sheet steric zipper, as are the Asn-X-Gln-X-Asn sequences studied here. Similarly, we expect that short segments of low complexity sequences can form steric zippers. The observation that polyamino acids form amyloid-like fibrils⁴² is consistent with the importance of side-chain interactions in steric zippers, notably size and shape complementarity.

The self-complementing GNNQQNY sequence is a segment of the yeast prion Sup35, a protein known to convert copies of itself to an amyloid fibril-like state. This fibrillar state has been shown to be at the basis of the transition to the [PSI⁺] prion state of Sup35 (refs 20, 21, 24, 25). Presumably, self-complementation by a steric zipper is a preliminary step in the process of molecular self-recognition that leads to conversion. Because the steric zipper involves nonspecific van der Waals forces, a given sequence may form more than one self-complementing steric zipper, possibly leading to amyloid fibril polymorphism and prion strains.

Regulation of protein concentration within cells and tissues takes on significance in preventing fibril formation in light of the structure-based arguments presented here that the standard free energy of fibril formation is not strongly negative. If, in fact, the dissolved and fibrillar forms of proteins are nearly iso-energetic in the biological milieu of an organism, there are two factors that influence the formation of amyloid-like fibrils. The first is the concentration of a protein in a given tissue. Breakdown in the cellular machinery that regulates protein synthesis or protein degradation could raise the concentration of protein monomers to the point of favouring an aggregated state. If the protein in question contains self-complementing segments of sequence, the aggregate could be the amyloid-like state. Chaperones that isolate proteins as they fold would be of critical importance when those sequences contain self-complementing segments. The second factor is the energetic barrier on the reaction pathway. The GNNQQNY structure suggests that several self-complementary segments must be properly arranged to act as a nucleus for fibril growth, presenting a significant barrier to fibril formation. However, once fibrils form at high protein concentration, the barrier to the reverse reaction—dissolution of the fibril—is even higher, rendering fibril formation difficult to reverse.

METHODS

Peptide crystallization. Lyophilized, synthetic GNNQQNY (AnaSpec) and NNQQNY (AnaSpec) peptides dissolve easily in water and aqueous solutions. Because of residual trifluoroacetic acid in the lyophilized peptide, dissolving the material in water results in a low pH solution; this low pH solution was used for crystallization.

GNNQQNY crystals were grown from a solution of 10 mg ml⁻¹ peptide in water (pH ~ 2.0) at about 20 °C. An orthorhombic crystal polymorph was previously grown^{6,7} from this condition. In later preparations, seed crystals from previous batches were used to promote faster and more reliable crystallization.

NNQQNY crystals were grown using the hanging-drop vapour diffusion method by mixing a 5:4:1 ratio of peptide solution, reservoir solution and additive solution, respectively. The peptide solution contained 30 mg ml⁻¹ NNQQNY in water. The reservoir solution contained 100 mM HEPES (pH 7.5) and 1 M sodium acetate. The additive solution contained 0.1 M zinc sulphate. The final pH of the drop was about 7.5, and crystals were grown at approximately 20 °C. GNNQQNY and NNQQNY crystals were transferred to a cryoprotectant (either 50% ethylene glycol/water or 50% glycerol/water) before data collection.

X-ray data collection and processing. X-ray diffraction data sets were collected

from the GNNQQNY and NNQQNY crystals at ESRF beamline ID13, equipped with a MAR CCD detector⁴³. Data were collected in 5° wedges at a wavelength of 0.975 Å using a 5-μm beam size. The crystals were cryo-cooled (100 K) for data collection. Owing to the extremely small focal size of the X-ray beam, the effect of localized radiation damage could be minimized by illuminating three different portions of the NNQQNY crystal during data collection (Fig. 1). All data were processed and reduced using Denzo/Scalepack from the HKL suite of programs⁴⁴.

Structure determination and refinement. An initial set of phases for the NNQQNY structure could be derived by the method of single wavelength anomalous dispersion (SAD) using the anomalous scattering signal from a well-ordered zinc ion. The location of the zinc ion was readily deduced from the presence of a 5-σ peak in an anomalous difference Patterson map (Supplementary Fig. 1). SAD phases were calculated with the program MLPHARE⁴⁵. Density modification with the program DM⁴⁵ significantly improved the interpretability of the electron density map, despite an extremely low solvent content (18%). A six-residue-long β-strand could be immediately recognized and modelled in the electron density with no ambiguity in orientation or position. Side-chain torsion angles were adjusted using the graphics program O⁴⁶. Coordinates were refined with the program REFMAC⁴⁷. Refinement statistics are reported in Table 1. The geometric quality of the model was assessed with the programs PROCHECK⁴⁸ and WHATIF⁴⁹. All residues were found in the most favoured region of the Ramachandran plot. The GNNQQNY structure could be refined by difference Fourier methods because its unit cell was nearly isomorphous with that of the NNQQNY crystal. Protein structures were illustrated using the program PyMOL⁵⁰.

Received 18 January; accepted 25 April 2005.

1. Sipe, J. D. & Cohen, A. S. Review: history of the amyloid fibril. *J. Struct. Biol.* **130**, 88–98 (2000).
2. Cohen, A. S. & Calkins, E. Electron microscopic observations on a fibrous component in amyloid of diverse origins. *Nature* **183**, 1202–1203 (1959).
3. Eanes, E. D. & Glenner, G. G. X-ray diffraction studies on amyloid filaments. *J. Histochem. Cytochem.* **16**, 673–677 (1968).
4. Geddes, A. J., Parker, K. D., Atkins, E. D. & Beighton, E. “Cross-β” conformation in proteins. *J. Mol. Biol.* **32**, 343–358 (1968).
5. Sunde, M. et al. Common core structure of amyloid fibrils by synchrotron X-ray diffraction. *J. Mol. Biol.* **273**, 729–739 (1997).
6. Balbirnie, M., Grothe, R. & Eisenberg, D. S. An amyloid-forming peptide from the yeast prion Sup35 reveals a dehydrated β-sheet structure for amyloid. *Proc. Natl Acad. Sci. USA* **98**, 2375–2380 (2001).
7. Diaz-Avalos, R. et al. Cross-β order and diversity in nanocrystals of an amyloid-forming peptide. *J. Mol. Biol.* **330**, 1165–1175 (2003).
8. Petkova, A. T. et al. Self-propagating, molecular-level polymorphism in Alzheimer's β-amyloid fibrils. *Science* **307**, 262–265 (2005).
9. Benzinger, T. L. et al. Propagating structure of Alzheimer's β-amyloid(10–35) is parallel β-sheet with residues in exact register. *Proc. Natl Acad. Sci. USA* **95**, 13407–13412 (1998).
10. Petkova, A. T. et al. A structural model for Alzheimer's β-amyloid fibrils based on experimental constraints from solid state NMR. *Proc. Natl Acad. Sci. USA* **99**, 16742–16747 (2002).
11. Jaroniec, C. P., MacPhee, C. E., Astrof, N. S., Dobson, C. M. & Griffin, R. G. Molecular conformation of a peptide fragment of transthyretin in an amyloid fibril. *Proc. Natl Acad. Sci. USA* **99**, 16748–16753 (2002).
12. Sunde, M. & Blake, C. C. From the globular to the fibrous state: protein structure and structural conversion in amyloid formation. *Q. Rev. Biophys.* **31**, 1–39 (1998).
13. Sumner Makin, O., Atkins, E., Sikorski, P., Johansson, J. & Serpell, L. C. Molecular basis for amyloid fibril formation and stability. *Proc. Natl Acad. Sci. USA* **102**, 315–320 (2005).
14. Serag, A. A., Altenbach, C., Gingery, M., Hubbell, W. L. & Yeates, T. O. Identification of a subunit interface in transthyretin amyloid fibrils: evidence for self-assembly from oligomeric building blocks. *Biochemistry* **40**, 9089–9096 (2001).
15. Torok, M. et al. Structural and dynamic features of Alzheimer's Aβ peptide in amyloid fibrils studied by site-directed spin labeling. *J. Biol. Chem.* **277**, 40810–40815 (2002).
16. Jimenez, J. L. et al. Cryo-electron microscopy structure of an SH3 amyloid fibril and model of the molecular packing. *EMBO J.* **18**, 815–821 (1999).
17. Kishimoto, A. et al. β-Helix is a likely core structure of yeast prion Sup35 amyloid fibers. *Biochem. Biophys. Res. Commun.* **315**, 739–745 (2004).
18. Williams, A. D. et al. Mapping Aβ amyloid fibril secondary structure using scanning proline mutagenesis. *J. Mol. Biol.* **335**, 833–842 (2004).
19. Wickner, R. B. [URE3] as an altered URE2 protein: evidence for a prion analog in *Saccharomyces cerevisiae*. *Science* **264**, 566–569 (1994).
20. Patino, M. M., Liu, J. J., Glover, J. R. & Lindquist, S. Support for the prion hypothesis for inheritance of a phenotypic trait in yeast. *Science* **273**, 622–626 (1996).

21. Serio, T. R. *et al.* Nucleated conformational conversion and the replication of conformational information by a prion determinant. *Science* **289**, 1317–1321 (2000).
22. King, C. Y. & Diaz-Avalos, R. Protein-only transmission of three yeast prion strains. *Nature* **428**, 319–323 (2004).
23. Tanaka, M., Chien, P., Nabar, N., Cooke, R. & Weissman, J. S. Conformational variations in an infectious protein determine prion strain differences. *Nature* **428**, 323–328 (2004).
24. DePace, A. H., Santoso, A., Hillner, P. & Weissman, J. S. A critical role for amino-terminal glutamine/asparagine repeats in the formation and propagation of a yeast prion. *Cell* **93**, 1241–1252 (1998).
25. Santoso, A., Chien, P., Osherovich, L. Z. & Weissman, J. S. Molecular basis of a yeast prion species barrier. *Cell* **100**, 277–288 (2000).
26. Jarrett, J. T. & Lansbury, P. T. Jr Seeding “one-dimensional crystallization” of amyloid: a pathogenic mechanism in Alzheimer’s disease and scrapie? *Cell* **73**, 1055–1058 (1993).
27. Lawrence, M. C. & Colman, P. M. Shape complementarity at protein/protein interfaces. *J. Mol. Biol.* **234**, 946–950 (1993).
28. Ivanova, M. I., Sawaya, M. R., Gingery, M., Attinger, A. & Eisenberg, D. An amyloid-forming segment of β 2-microglobulin suggests a molecular model for the fibril. *Proc. Natl Acad. Sci. USA* **101**, 10584–10589 (2004).
29. Jimenez, J. L. *et al.* The protofibril structure of insulin amyloid fibrils. *Proc. Natl Acad. Sci. USA* **99**, 9196–9201 (2002).
30. Perutz, M. F., Johnson, T., Suzuki, M. & Finch, J. T. Glutamine repeats as polar zippers: their possible role in inherited neurodegenerative diseases. *Proc. Natl Acad. Sci. USA* **91**, 5355–5358 (1994).
31. Pickersgill, R. W. A primordial structure underlying amyloid. *Structure (Camb.)* **11**, 137–138 (2003).
32. Wetzel, R. Ideas of order for amyloid fibril structure. *Structure (Camb.)* **10**, 1031–1036 (2002).
33. Perutz, M. F., Finch, J. T., Berriman, J. & Lesk, A. Amyloid fibers are water-filled nanotubes. *Proc. Natl Acad. Sci. USA* **99**, 5591–5595 (2002).
34. Govaerts, C., Wille, H., Prusiner, S. B. & Cohen, F. E. Evidence for assembly of prions with left-handed β -helices into trimers. *Proc. Natl Acad. Sci. USA* **101**, 8342–8347 (2004).
35. Varley, P. *et al.* Kinetics of folding of the all- β sheet protein interleukin-1 beta. *Science* **260**, 1110–1113 (1993).
36. Sivaraman, T., Kumar, T. K., Chang, D. K., Lin, W. Y. & Yu, C. Events in the kinetic folding pathway of a small, all β -sheet protein. *J. Biol. Chem.* **273**, 10181–10189 (1998).
37. Eisenberg, D., Wesson, M. & Yamashita, M. Interpretation of protein folding and binding with atomic solvation parameters. *Chem. Scr.* **29A**, 217–221 (1989).
38. Coulson, C. A. & Eisenberg, D. Interactions of H₂O molecules in ice. *Proc. R. Soc.* **291**, 445–453 (1966).
39. Richardson, J. S. & Richardson, D. C. Natural β -sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc. Natl Acad. Sci. USA* **99**, 2754–2759 (2002).
40. Lopez de la Paz, M. & Serrano, L. Sequence determinants of amyloid fibril formation. *Proc. Natl Acad. Sci. USA* **101**, 87–92 (2004).
41. Tjernberg, L., Hosia, W., Bark, N., Thyberg, J. & Johansson, J. Charge attraction and beta propensity are necessary for amyloid fibril formation from tetrapeptides. *J. Biol. Chem.* **277**, 43243–43246 (2002).
42. Fandrich, M. & Dobson, C. M. The behaviour of polyamino acids reveals an inverse side chain effect in amyloid structure formation. *EMBO J.* **21**, 5682–5690 (2002).
43. Riek, C. Recent developments in micro-diffraction on protein crystals. *J. Synchrotron Radiat.* **11**, 4–6 (2004).
44. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
45. Collaborative Computational Project Number 4, The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
46. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
47. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53**, 240–255 (1997).
48. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK – a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283–291 (1993).
49. Vriend, G. & Sander, C. Quality control of protein models: directional atomic contact analysis. *J. Appl. Crystallogr.* **26**, 47–60 (1993).
50. DeLano, W. L. *The PyMOL User’s Manual* (DeLano Scientific, San Carlos, California, 2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the late Carl Branden for initiating the UCLA–ESRF collaboration; D. L. D. Caspar, R. Diaz-Avalos, Y. Fujiyoshi, R. G. Griffin, S. Larsen, K. Mitsuoka, P. W. Stevens, J.-H. Her and T. O. Yeates for discussions; S. Horvath for peptide synthesis; and NIH, NSF, HHMI and USPHS National Research Service Award for support.

Author Information The structures of GNNQQNY and NNQQNY have been deposited in the Protein Data Bank with accession codes 1yjp and 1yjo, respectively. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.E. (david@mbi.ucla.edu).

The structure of the myosin VI motor reveals the mechanism of directionality reversal

Julie Ménétrey^{1*}, Amel Bahloul^{1*}, Amber L. Wells², Christopher M. Yengo², Carl A. Morris², H. Lee Sweeney² & Anne Houdusse¹

Here we solve a 2.4-Å structure of a truncated version of the reverse-direction myosin motor, myosin VI, that contains the motor domain and binding sites for two calmodulin molecules. The structure reveals only minor differences in the motor domain from that in plus-end directed myosins, with the exception of two unique inserts. The first is near the nucleotide-binding pocket and alters the rates of nucleotide association and dissociation. The second unique insert forms an integral part of the myosin VI converter domain along with a calmodulin bound to a novel target motif within the insert. This serves to redirect the effective 'lever arm' of myosin VI, which includes a second calmodulin bound to an 'IQ motif', towards the pointed (minus) end of the actin filament. This repositioning largely accounts for the reverse directionality of this class of myosin motors. We propose a model incorporating a kinesin-like uncoupling/docking mechanism to provide a full explanation of the movements of myosin VI.

The myosin superfamily is composed of 18 classes of molecular motor proteins, almost all of which traffic towards the barbed (plus) end of actin filaments¹. Class VI myosins were the first of the superfamily identified as trafficking towards the pointed (minus) end of the actin filament². They function in several critical intracellular processes such as vesicular membrane traffic, cell migration, maintenance of stereocilia, and mitosis^{3–6}.

The current view of how myosin motors couple ATP hydrolysis and actin binding to movement is known as the lever-arm hypothesis⁷. In essence the proposed mechanism is that nucleotide binding, hydrolysis and product release are all coupled to small movements within the myosin motor core. These movements are amplified and transmitted through a region that has been termed the 'converter' domain to a lever arm consisting of a target helix and associated light chains and/or calmodulin (CaM) molecules. The lever arm further amplifies the motions of the converter domain into large directed movements. Consistent with the lever-arm hypothesis is the observation that the stroke size is proportional to the length of the lever arm^{8–10}. In the absence of actin, ATP hydrolysis occurs but product release is slow, thus trapping the lever arm in a primed or pre-powerstroke position. Binding to actin causes the release of products, a plus-end-directed movement of the lever arm, and the generation of force concomitant with the formation of strong binding between myosin and actin.

Perhaps because of their reverse directionality, myosin VI motors have several additional unusual features. First, the motor domain itself contains two inserts that are unique within the myosin superfamily. The first is near the nucleotide-binding pocket; the second is between the converter and the IQ motif. (These correspond to residues Cys 278–Ala 303 and Pro 774–Tyr 812, respectively, and are called insert 1 and insert 2 hereafter.) Single molecules of dimeric myosin VI, like myosin V, are capable of taking multiple steps

(processive movement) of 30–36 nm by means of a hand-over-hand mechanism along actin filaments^{11,12}. This large step size is surprising within the context of the lever-arm hypothesis, in that myosin VI binds only two CaM molecules per head¹³, whereas myosin V binds six. Furthermore, a construct truncated after the binding site for the second CaM (the IQ motif) has a stroke (non-processive displacement after an actin encounter) of 12 nm, less than half the step size of the two-headed molecule¹⁴. Thus, the true composition of the myosin VI 'lever arm' is unclear.

The fact that myosin VI binds two CaM molecules per head was surprising because there is only one conventional CaM-binding site (IQ motif) per head¹³. It was postulated initially that the unique insert 2 forms part of the 'converter' of myosin VI and redirects the lever arm (that is, IQ-bound CaM) towards the minus end of an actin filament². That hypothesis was brought into question by studies on chimaeric molecules of myosin V and VI, in which the removal or addition of insert 2 did not alter directionality¹⁵. The finding that insert 2 provides a second CaM-binding site¹³ raises the possibility that its purpose could be simply to lengthen the lever arm. To gain insight into the function of these two unique inserts within the myosin VI motor, we determined the crystal structure of two fragments of myosin VI.

Actin-binding cleft of nucleotide-free myosin VI

We expressed CaM together with pig myosin VI constructs coding for either the motor domain (containing inserts 1 and 2) alone (residues 1–816, hereafter called MD^{ins2}) or the motor domain and the helical CaM-binding domain (IQ motif) (residues 1–859, hereafter called MD^{ins2}IQ) in *Spodoptera frugiperda* (Sf9) cells. Crystals were obtained in the absence of nucleotide and diffracted to 2.4 and 2.9 Å for the MD^{ins2} and MD^{ins2}IQ constructs, respectively. The two refined structures obtained from different crystal forms superimpose

¹Structural Motility, Institut Curie CNRS, UMR144, 26 rue d'Ulm, 75248 Paris cedex 05, France. ²Department of Physiology, University of Pennsylvania School of Medicine, 3700 Hamilton Walk, Philadelphia, Pennsylvania 19104-6085, USA.

*These authors contributed equally to this work.

very well with a root-mean-square (r.m.s.) deviation of 0.694 Å for 787 C α atoms of the motor domain (Gly 4–Tyr 812). These structures therefore define the nucleotide-free state of myosin VI, which differs greatly from that found for plus-end myosin motors in the reversed orientation of the lever arm (Fig. 1). However, the overall conformation of the myosin VI motor domain is similar to that recently described for myosin V¹⁶ and for the *Dictyostelium discoideum* myosin II motor in the absence of nucleotide¹⁷: the twist of the central β -sheet in all these structures differs greatly from that found for the post-rigor (ATP) and pre-powerstroke (ADP.P_i) states of the motor^{16–18}. The myosin V structure has been suggested to be in a ‘rigor-like’ state, meaning that it is the state that myosin populates when bound to actin in the absence of nucleotide^{16,19}. Several differences—minor in amplitude but not in significance—distinguish the myosin VI and II¹⁷ nucleotide-free structures from the myosin V rigor-like state. The major cleft in the molecule that closes on binding to actin is not completely closed in myosin VI, in contrast with myosin V (Supplementary movie S1). Moreover, the set of interactions found between the nucleotide-binding elements in the rigor-like state¹⁸ are also not found in myosin VI, where differences in the distortion of the central β -sheet (Supplementary movie S2) are seen. The kinetics of its interaction with actin²⁰ also demonstrates that, in the absence of nucleotide and actin, myosin VI does not populate the rigor-like state. However, docking of the nucleotide-free myosin VI structure

into reconstructions from electron microscopy images of the actomyosin VI rigor complex² confirms that the state that we crystallized is close to the conformation of myosin VI at the end of its power stroke (Fig. 1a).

Insert 1 modulates nucleotide binding and release

Myosin VI not only moves in the reverse direction but has unusual kinetic properties. Most notable is the slow rate (40-fold slower than myosin V) of ATP-induced release of myosin VI from the actomyosin complex^{20,21}. This is primarily due to a weak affinity for ATP, which is weaker than that of either myosin VI or actomyosin VI for ADP (see Supplementary Table S1). The first of the two unique inserts in myosin VI (Cys 278–Ala 303) belongs to the U50kDa subdomain and is located close to the nucleotide-binding site, near switch I (Figs 1 and 2a). Switch I has a critical function in binding both the Mg²⁺ ion and the γ -phosphate of ATP in the active site. It is now believed that sequential conformational rearrangements of switch I are essential in controlling nucleotide release from and binding to the motor^{22,23}. The location of insert 1 leads to the prediction that it might have evolved to provide unique kinetic characteristics that are potentially important for a reverse-directed motor.

To test the influence of insert 1 on the kinetics of this reverse-direction motor, we have engineered a myosin VI lacking insert 1 (Δ C278–A303) and measured the impact on the actomyosin ATPase activity and nucleotide binding and release. These results show

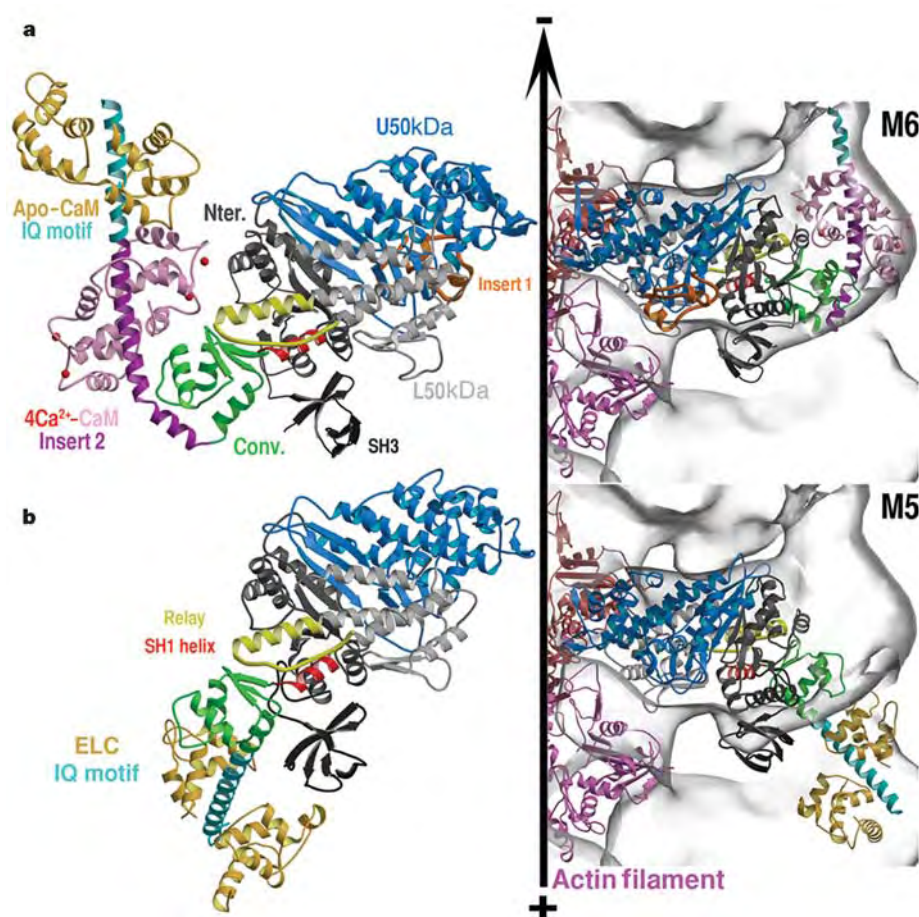


Figure 1 | Crystal structure of myosin VI. **a**, A ribbon diagram representation of the nucleotide-free myosin VI structure highlights its two specific inserts, the four subdomains (N-terminal (Nter.), U50kDa, L50kDa and converter (conv.)) of the motor domain and two connectors (relay and SH1 helix) that control the converter. Insert 2 recruits 4Ca²⁺-CaM; the IQ motif interacts with apo-CaM. **b**, Comparison with the myosin V rigor-like

state¹⁶. Note the similarity in the overall conformation of the motor domains and the striking difference in the orientation of the lever arm (composed of an IQ motif and an associated light chain). On the right are the same structures rotated by 180° along the actin filament axis (black arrow), obtained by fitting the three-dimensional map of the actin–myosin VI rigor complex (obtained by cryo-electron microscopy)².

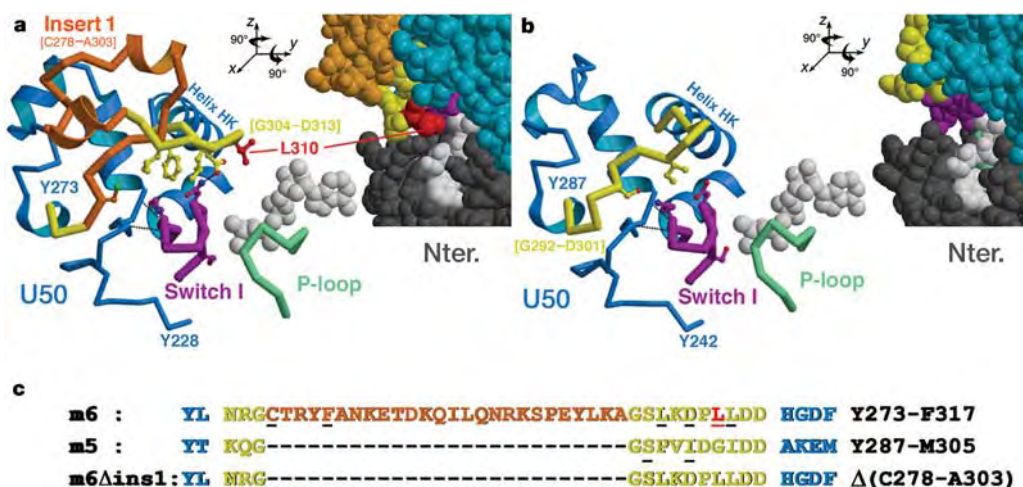


Figure 2 | Insert 1 modulates nucleotide binding and switch I flexibility.
a, Insert 1 is found at the surface of the U50kDa subdomain of nucleotide-free myosin VI near the nucleotide-binding site. **b**, The same region is shown for the myosin V rigor-like state, which lacks this insert. Note the conformational change in the loop (yellow) in which insert 1 is inserted. Residues that interact with switch I are shown in ball-and-stick form, and

that insert 1 has a crucial function in slowing ADP release and ATP-induced actomyosin dissociation (see data in Supplementary Table S1). The myosin VI structure reveals that the presence of insert 1 does not alter the conformation of switch I relative to the U50kDa subdomain, which is the same in all myosin states determined so far (Fig. 2). In contrast, the small loop (Gly 304–Asp 313) that follows insert 1 is drastically repositioned and interacts strongly with switch I. This loop also protrudes within the nucleotide-binding

ATP molecules are modelled in white CPK atom representation. In addition, in the upper-right corner, a surface CPK⁴¹ representation of this region is depicted in a different orientation, to evaluate the accessibility of the nucleotide-binding pocket for the ATP molecule. **c**, Sequence alignment of myosin V and myosin VI near insert 1, coloured as in **a** and **b**. Residues involved in interacting with switch I are underlined.

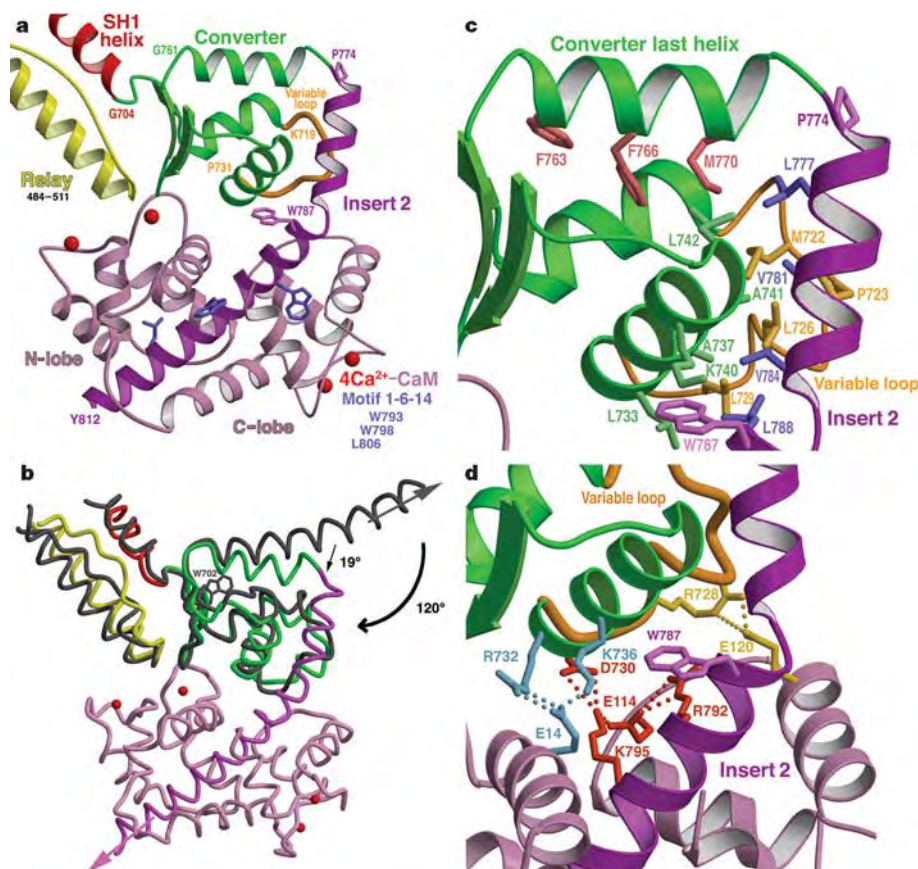


Figure 3 | Reorientation of the myosin VI lever arm by its unique converter. **a**, The proximal part of insert 2 (P774–W787; purple/grey) interacts on the surface of the converter, in particular with a loop whose sequence is highly variable between myosin classes but quite conserved within the myosin VI class. The distal part of insert2 (W787–Y812; purple/dark-pink) recruits 4Ca^{2+} -CaM. **b**, In comparison with the rigor-like state of chicken myosin V (black), the linkages found between the relay/SH1 helix connectors and the converter β -sheet are conserved but the orientation of the last helix of the converter (G761–D773) differs by 19° . Note that W702 of myosin V (A709 in myosin VI) would interfere with this orientation. **c**, Close-up view of the apolar interactions that stabilize the proximal part of insert 2 on the surface of the converter. Note also how three anchoring hydrophobic side chains (light red) stabilize the orientation of the last helix. **d**, Close-up view of the salt-bridge interactions between the converter and both lobes of CaM.

the U50kDa subdomain, which would explain its influence on both ADP release and ATP-induced dissociation of the motor from actin at saturating ATP concentrations (k_{+2}) (see Supplementary Information).

Insert 2 redirects the lever arm

Functional studies of chimaeric molecules between myosin V and VI were interpreted as evidence that the motor domain, rather than insert 2, provides the major determinants for directionality reversal¹⁵. However, the nucleotide-free myosin VI structure shows that its motor domain is very similar to that of myosin V, including the position of the converter, the specific subdomain designed to direct lever arm movement in myosin motors. The converter position is controlled by the specific conformation of two connectors: the relay and the SH1 helix (Figs 1 and 3a, b). As found in all plus-end myosins, the set of hydrophobic interactions that maintain the relay and the converter closely linked to one another (in all states of the motor) are also conserved in myosin VI. However, there are sequence variations specific for myosin VI that are clustered in the interface between these two connectors (Supplementary Movie S3). Their effect in modulating the precise orientation of the converter is minimal in the nucleotide-free state of myosin VI, but they could be critical in other states of the cycle, such as the pre-powerstroke state.

The truly unique feature of the two myosin VI structures is the reversal of the lever arm direction by insert 2 (Pro 774–Tyr 812; Fig. 3). The proximal part of the insert (Pro 774–Trp 787) wraps around the converter, rather than emerging as a straight helix from the converter, and the distal part of the insert (Trp 787–Tyr 812) forms a previously unseen CaM-binding motif (see Fig. 4). Both the insert and its associated CaM molecule (with four bound Ca^{2+} ions) make specific interactions with the converter, many involving a variable loop (Lys 719–Pro 731); Fig. 3a. The net result of these interactions is that the IQ helix emerges about 120° from the position that the IQ helix would in any other myosin (Fig. 3b). This redirects the IQ helix and its bound CaM (which form a 'lever arm') towards the minus end of the actin filament (Fig. 1).

Unique structural differences of the myosin VI converter are crucial for its close interaction with insert 2 and thus for the reorientation of the lever arm. In particular, the orientation of the last helix of the myosin VI converter differs by about 19° from that found in plus-end myosins when the converters are superimposed (Fig. 3b). This helix is well anchored against the rest of the converter through several large hydrophobic side chains. It ends with a proline residue (Pro 774), which favours a 90° turn at the beginning of insert 2, promoting its wrapping around the converter. The variable loop of the converter and the helix that follows create a small hydrophobic cavity in which small hydrophobic side chains of insert 2 fit (Fig. 3c). A break in the helix of insert 2 at position Val 784 further extends the surface of interaction with the converter (through Trp 787 and Leu 788) and promotes interactions between the CaM held by the distal part of this insert and the converter—in particular by means of six salt bridges (Fig. 3d). In contrast, the helix remains straight between insert 2 and the IQ motif, and the stiffness at the MD/lever-arm junction in this myosin is probably similar to that found for other myosins. The functional converter of myosin VI that redirects the lever arm in the opposite direction in the rigor state is therefore composed of the normal converter plus insert-2–CaM.

Insert 2 as a novel CaM-binding motif

Although the sequence of the myosin VI-specific insert 2 does not correspond to any CaM-binding motif described so far²⁴, we previously reported that it recruits 4Ca^{2+} -CaM and that dissociation of the Ca^{2+} ions cannot occur and regulate this interaction under physiological conditions¹³. Consistent with these observations, the myosin VI structure reveals that the distal part of insert 2 contains a novel CaM-binding motif (Fig. 4). Following the precedent for other

targets of CaM^{25} , such as the 1-8-14 motif found in the smooth myosin light chain kinase, we name this motif 1-6-14 based on the distances between three key hydrophobic residues (Trp 793, Trp 798 and Leu 806) buried in the interaction with CaM. In comparison with other Ca^{2+} -CaM–target (1-8-14) complexes²⁵, the polarity of

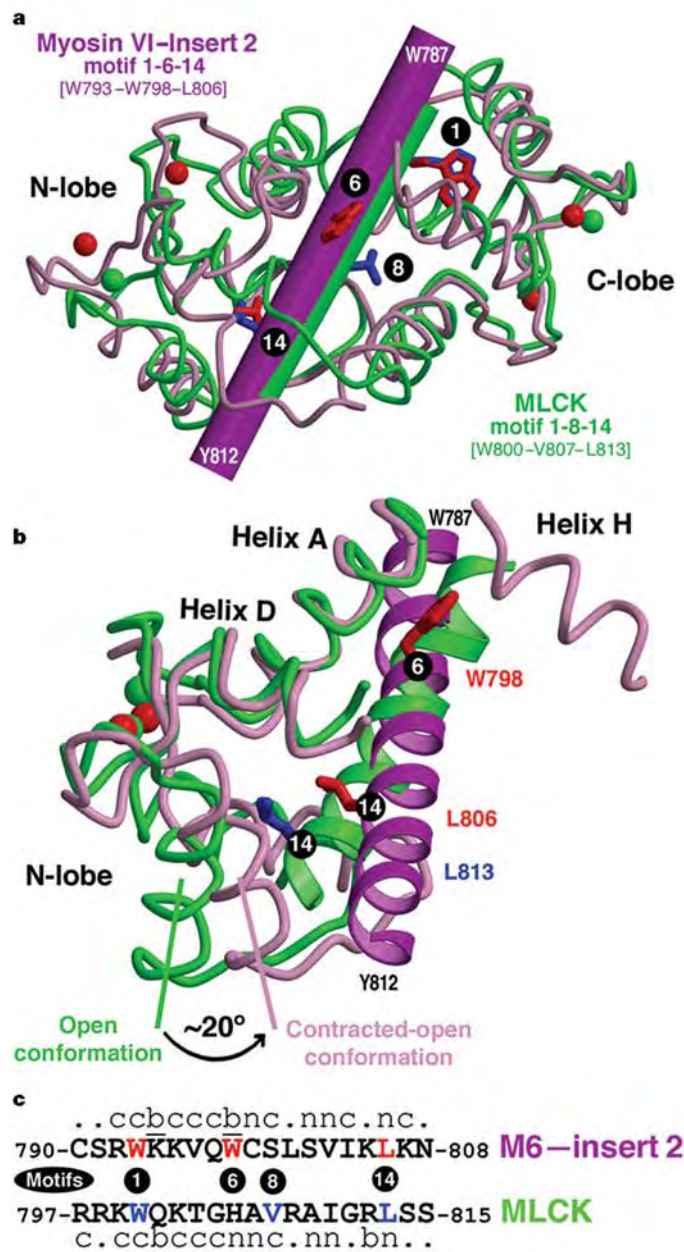


Figure 4 | A new CaM-binding motif that interacts strongly with 4Ca^{2+} -CaM. **a**, The overall conformation and polarity of the insert-2–CaM complex (new 1-6-14 motif) is compared with those observed when CaM interacts with myosin light chain kinase (MLCK) (classic 1-8-14 motif) (target peptides superimposed). Note that in both cases the C-lobe of CaM in an open conformation grips the N-terminal region of the target sequence, largely through the first anchoring hydrophobic residue (W793/W800). **b**, In contrast, comparison of the N-lobes (helices A and D superimposed) shows differences in their conformation (closure differs by 20°) and in the target peptide position (note the 14th anchoring residue position) within the lobe. Note that the sixth anchoring residue of the 1-6-14 motif (W798) interacts strongly with both lobes of CaM (helices A and H). **c**, Sequence comparison of the two CaM-binding motifs. The letters n, c and b indicate whether each residue of these motifs interacts with the N-lobe, the C-lobe or both lobes of CaM, respectively.

this Ca^{2+} -CaM-insert 2 complex is similar: the C-lobe of CaM grips the amino-terminal region of the target sequence (Fig. 4a). Whereas the C-lobe of CaM adopts a classic open conformation, that found for the N-lobe differs from any other complex described so far in being more closed by about 20° (Fig. 4b). (It is in fact close to that recently described as a contracted-open conformation²⁶.) The target helix is therefore found in a much less buried position within this lobe than in a classic open lobe (Fig. 4b). After close inspection of this structure, the selection of such a conformation for the N-lobe seems to be induced by the target motif itself, not the surface contacts that this lobe makes with the rest of the motor domain. In particular, the presence of a large hydrophobic side chain in the sixth position, Trp 798, has a critical function because it interacts strongly with helix A and the last helix of the C-lobe (Fig. 4b). By providing strong hydrophobic interactions and by selecting for relative orientations of

the two lobes of CaM that allow them both to interact with the converter, this motif is remarkably well suited to impose a structural role on the recruited CaM (Fig. 4c).

Discussion

What is at first striking about myosin VI is that reverse directionality has been accomplished with only minor changes within the motor domain itself. The first unique insert, insert 1, is not positioned to affect directionality by altering the structure of the motor domain. However, it does have a major effect on kinetic properties by slowing ADP release and creating a long-lived rigor state. Insert 1 is therefore essential for the processivity of the motor and would be critical for the anchoring role of myosin VI, which is dependent on slow ATP binding relative to ADP binding²⁷.

The second unique insert, insert 2, is what initially called attention to myosin VI² and indeed is the most interesting feature of this structure. It reveals that myosin VI has tightly coupled both this insert and a CaM to the conventional myosin converter to create a unique converter subdomain. The purpose of this design is to redirect the lever arm towards the minus end of the actin filament. This undoubtedly is essential for reverse directionality. For a single-headed myosin, it would reposition the lever arm towards the minus end of the actin filament in the rigor state (end of power stroke), and for a two-headed processive myosin it would also bias the unbound lead head towards minus-end-directed binding sites.

It has been shown²⁸ that repositioning the lever arm by 180° , rather than 120° as for myosin VI, is sufficient to reverse the directionality using a single-headed plus-end myosin motor domain. To assess whether the repositioning of the lever arm is sufficient to explain the myosin VI stroke, a model for myosin VI at the beginning of the power stroke was obtained by using the pre-powerstroke state of plus-end motors (see Methods). A minus-end-directed movement (that is, reversal) is indeed produced but the component of the displacement parallel to the actin filament would be only about 2.5 nm (Fig. 5a). This is one-fifth of the 12-nm stroke size measured from optical trap studies for a MD^{ins2}IQ molecule¹⁴. Similar modelling for a truncated myosin V motor would predict a stroke of about 7 nm (Fig. 5b), which is what was measured in optical trap studies⁶. Note that for either myosin V or this myosin VI model, the rotation of the converter contributes a plus-end-directed stroke. Although this contribution adds to the lever arm displacement for a plus-end-directed motor, it must be overcome by the lever arm contribution (for example, by increasing lever arm length) for a minus-end-directed motor. This is illustrated in a model of the artificial lever

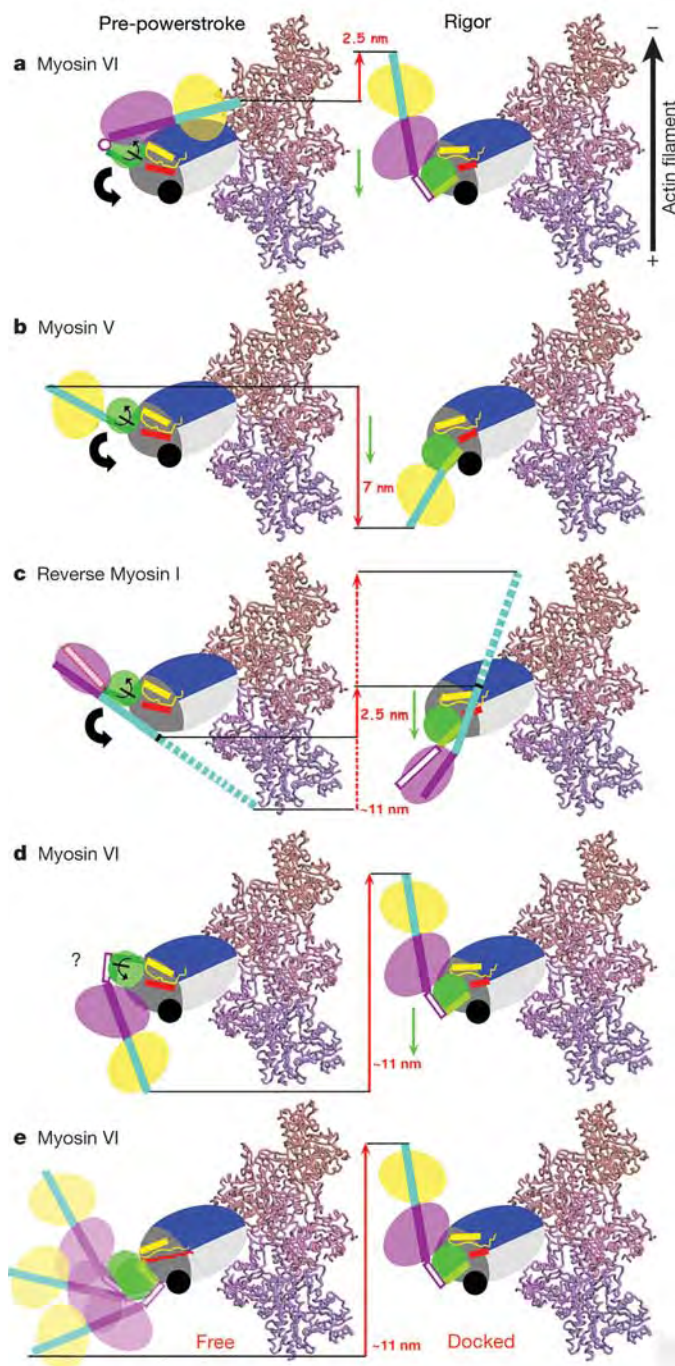


Figure 5 | Directionality of movement and power stroke in myosin motors.

a, b, Schematic drawings of the myosin VI (**a**) and myosin V (**b**) structural models (see Methods) before and after force generation. Similar colours to those in Fig. 1 are used. Note in particular how conformational changes in the relay (yellow) and SH1 helix (red) lead to the rotation (black arrow) of the converter (green). The red arrow represents the predicted F-actin displacement (stroke) for these models; the green arrow indicates the converter contribution for this stroke. **c,** For reverse myosin I, the solid arrow indicates the stroke that would be produced with a lever arm of about 4 nm (that equivalent to one IQ motif) and the dotted arrow corresponds to the stroke generated by an approximately 14-nm lever arm as described for this engineered motor²⁸. **d, e,** Two mechanisms could account for the ~12-nm stroke of the myosin VI MD^{ins2}IQ. If the converter remains coupled to the motor domain (**d**), it must adopt an orientation that differs by about 90° from that found for plus-end motors in the pre-powerstroke state. Alternatively, unwinding of the SH1 helix in the weak actin-binding states would decouple the converter from the motor domain (**e**). In this case, the relay-converter interactions would be maintained but the relay helix would not be bent in the pre-powerstroke state because steric clashes with the SH1 helix are eliminated. Thus, the converter would be biased towards the plus end of the actin filament. Recoupling of the converter to the motor domain would occur on strong binding to actin.

arm that achieved reversal of myosin I²⁸ using a plus-end converter rotation and a lever arm 14 nm long (Fig. 5c).

Because a MD^{ins2}IQ myosin VI molecule has an inexplicably large¹⁴ minus-end-directed stroke² with a very short (4-nm) lever arm, we conclude that myosin VI has a quite different pre-power-stroke structure from other myosins. This could be accomplished in at least two fundamentally different ways. First, the position of the converter could differ from that of other characterized myosins in all of the nucleotide states of myosin VI. There is already evidence from cryo-electron microscopy that the rotation from the actomyosin ADP to the rigor state is in a different direction². Because the rearrangements necessary for this transition are of small amplitude, it is possible to imagine that they differ in myosin VI because of the small alterations in the nature of the β -sheet and SH1 helix (Supplementary movies S2 and S3). However, major rearrangements occur upon ATP binding to generate the post-rigor and pre-power-stroke states¹⁸, and it is difficult to imagine how these subtle motor domain differences could cause the nearly 90° change in rotation of the converter in the pre-powerstroke state that would be necessary to generate the large step size of the MD^{ins2}IQ molecule¹⁴ (see Fig. 5d). Such an altered rotation would probably require a major redesign of the motor domain, which the structure of myosin VI clearly reveals did not occur.

The second and more easily envisaged possibility is that myosin VI has not truly reversed the normal myosin power stroke but has essentially abolished it and evolved a mechanism similar to that proposed for 'conventional' kinesin^{29,30}. It is thought that movement in kinesin is accomplished by a reversible, nucleotide-state-dependent docking and undocking of its neck-linker region. In the two-headed kinesin molecule, intramolecular strain gates the docking and undocking, and the docked head biases the diffusive search for a new tubulin-binding site of the unbound head during processive movement. To apply this mechanism to myosin VI, either the insert-2-CaM or the entire converter subdomain must uncouple from the motor domain in the pre-powerstroke state. Insert-2-CaM seems to be an integral part of the converter, and as can be seen in the pre-powerstroke model (Fig. 5a) there would be no steric hindrance between this unique myosin VI converter and the motor domain to drive such an uncoupling. However, we propose that uncoupling of the entire converter from the motor domain could be induced by an SH1 helix unwinding in the post-rigor and pre-powerstroke states (Fig. 5e). A cluster of sequence differences specific for myosin VI are indeed found in the cavity of the SH1 helix (Supplementary movie S3) and there is structural evidence for unwinding of this connector in myosin II after nucleotide binding³¹. Thus, in myosin VI, the SH1 helix would be roughly analogous to the neck-linker of kinesin.

Consistent with this hypothesis are data that placed fluorescent probes either on the IQ-CaM or on the N-terminal subdomain of two-headed myosin VI³². By using FIONA (for fluorescence imaging with one nanometer accuracy), large fluctuations in the position of the IQ-CaM, but not in the position of the motor domain, were observed during processive movement. These fluctuations disappeared when ATP was removed and the heads were strongly bound to actin with either ADP or no nucleotide³². Thus, on the basis of these observations and our new structural insights, we propose that myosin VI (M.ADP.P_i) binds to actin with an uncoupled converter domain. In the absence of strain, strong binding to actin accompanies cleft closure by means of the central β -sheet distortion. These rearrangements would lead to an interface between the SH1 helix and the N-terminal subdomain that would favour recoupling of the converter to the motor domain in the strong actin-binding actomyosin ADP state. Recoupling of the lead head for a two-headed processive myosin would be prevented, or greatly slowed, until the rear head detached. This is consistent with both the FIONA study³² and direct measurements of pyrene-actin quenching (strong binding) with the myosin VI dimer in the presence of ATP³³. When ADP dissociates, further cleft closure would lead to a small rotation of the

lever arm in the minus-end direction². The net displacement on actin (stroke) would be the difference between the average position of the lever arm when uncoupled and the stable rigor position.

This uncoupling model could account for movement of about 6–7 nm of the MD^{ins2}IQ construct, assuming that there is no positional bias of the uncoupled converter (Fig. 5e). However, as depicted in this model, any biasing towards the plus end of the actin filament would increase the stroke and step size. This biasing might come from the relay, which, on the basis of the myosin II structure with an unwound SH1 helix³¹, would maintain connections with the converter and would be plus-end directed because of the loss of the steric hindrance that normally bends the relay helix in the minus-end direction in the pre-powerstroke state^{7,31,34}. Additional experiments and structures will be necessary to test this proposed mechanism of directionality reversal. Undoubtedly, this unique myosin family member has yet more surprises to reveal.

METHODS

Crystallization and data collection. Protein expression and purification have been described previously^{2,13} and are detailed in Supplementary Methods. Crystals of myosin VI were first obtained with the short MD^{ins2}IQ [1–839] construct by vapour diffusion. Spontaneous nucleation occurred at 4 °C in hanging drops with equal amounts of reservoir solution (containing 8–10% PEG 8000, 50 mM Tris pH 8.0, 3% propan-2-ol and 3% tert-butanol) and stock solution of the myosin VI MD^{ins2}IQ at 10–12 mg ml⁻¹. Crystals of the MD^{ins2} [2–816] and long MD^{ins2}IQ [1–859] constructs were initially obtained by cross-seeding with crystals of the short MD^{ins2}IQ and then improved by seeding in solution containing 8–10% PEG 8000, 50 mM MES pH 6.7, 150 mM NH₄SO₄, 3% propan-2-ol and 3% tert-butanol. Before freezing and data collection, the crystals were transferred stepwise into a final cryoprotectant solution containing 16% PEG 8000 and 25% glycerol. X-ray data sets were collected at the European Synchrotron Radiation Facility beamlines at 100 K. The MD^{ins2} construct was crystallized in a monoclinic crystal form, and both MD^{ins2}IQ constructs were crystallized in an orthorhombic one. The long MD^{ins2}IQ crystals are similar to those of the short MD^{ins2}IQ except for the *b* axis whose length accommodates the difference in lever arm length. The short MD^{ins2}IQ, long MD^{ins2}IQ and MD^{ins2} crystals diffracted to 3.5, 2.9 and 2.4 Å resolution, respectively. All data sets were integrated and scaled with either the HKL package³⁵ or the CCP4 suite³⁶ (see Table 1 and Supplementary Table S2 for statistics on the data collection).

Structure determination and refinement. The myosin VI structure was solved by molecular replacement with the myosin V MDE model (PDB code 1OE9) by using the program AMoRe³⁷ with data at 3.5 Å resolution from the short MD^{ins2}IQ or the MD^{ins2} crystals. Several steps of rigid body fitting were performed with AMoRe (each subdomain was considered as a rigid group). Model building and refinement of the motor domain, insert 2 and its 4Ca²⁺-bound CaM were performed at 2.4 Å resolution with data from the MD^{ins2} crystals with Turbo³⁸, and CNS1.1 (ref. 39) and Refmac5 (CCP4 suite)³⁶. The long MD^{ins2}IQ structure was solved by molecular replacement³⁷ by using the coordinates from this 2.4-Å-refined myosin VI MD^{ins2} structure. Clear density was easily interpreted for the N-terminal part of the IQ helix and most of the carboxy-terminal lobe of its CaM, which were modelled as polyalanines. The C-terminal part of the IQ motif and most of the N-lobe of its CaM are not stabilized by crystal packing interactions. They have a high level of flexibility in

Table 1 | X-ray statistics on data collection and refinement

Parameter	Long MD ^{ins2} IQ	MD ^{ins2}
Resolution (Å)	2.9	2.4
<i>R</i> _{work} / <i>R</i> _{free}	26.3/30.3	20.6/25.4
Number of atoms		
Protein	7,991	7,403
Ligand/ion	5/4	36/4
Water	49	285
<i>B</i> -factors		
Protein	46.01	48.76
Ligand/ion	35.2/79.4	60.9/79.7
Water	28.37	44.90
r.m.s. deviations		
Bond lengths (Å)	0.013	0.011
Bond angles (°)	1.314	1.326

the crystal and were not included in the refined model (see Supplementary Fig. S1). Water molecules were placed with Arp/Warp program⁴⁰. Crystallographic statistics are summarized in Table 1. Note that all diagrams for the figures and the movies were computed using MOLSCRIPT/Raster3D⁴¹.

Pre-powerstroke state modelling. Models of the myosin VI and V pre-powerstroke states were obtained from known myosin II pre-powerstroke structures. The scallop striated muscle myosin II pre-powerstroke structure⁴² (PDB code 1DFL) provided coordinates for the motor domain. Positions for the myosin VI and V converter/lever arms were obtained by superimposing their converter on that of the myosin II converter. Similarly, the lever arm for the engineered reverse myosin I motor was obtained from a molecular model²⁸ (D. Manstein, personal communication). Note that this modelled myosin VI pre-powerstroke state assumes that the converter and the insert-2–CaM interactions are not broken during the catalytic cycle, which seems reasonable in view of their strong interactions. Note that no steric clash is generated between insert-2–CaM and the motor domain when this myosin II converter rotation is applied to myosin VI. The pre-powerstroke and nucleotide-free structural models (PDB codes 1OE9 and 2BKI) were docked on F-actin as described previously¹⁹ and a schematic drawing is shown for clarity in Fig. 5. The stroke resulting from this model was measured as the component of the displacement parallel to the actin filament.

Received 1 February; accepted 1 April 2005.

- Berg, J. S., Powell, B. C. & Cheney, R. E. A millennial myosin census. *Mol. Biol. Cell* **12**, 780–794 (2001).
- Wells, A. L. *et al.* Myosin VI is an actin-based motor that moves backwards. *Nature* **401**, 505–508 (1999).
- Geisbrecht, E. R. & Montell, D. J. Myosin VI is required for E-cadherin-mediated border cell migration. *Nature Cell Biol.* **4**, 616–620 (2002).
- Hasson, T. Myosin-VI: two distinct roles in endocytosis. *J. Cell Sci.* **116**, 3453–3461 (2003).
- Buss, F., Spudich, G. & Kendrick-Jones, J. Myosin VI: Cellular functions and motor properties. *Annu. Rev. Cell Dev. Biol.* **20**, 649–676 (2004).
- Millo, H., Leaper, K., Lazou, V. & Bownes, M. Myosin VI plays a role in cell-cell adhesion during epithelial morphogenesis. *Mech. Dev.* **121**, 1335–1351 (2004).
- Geeves, M. A. & Holmes, K. C. Structural mechanism of muscle contraction. *Annu. Rev. Biochem.* **68**, 687–728 (1999).
- Purcell, T. J., Morris, C., Spudich, J. A. & Sweeney, H. L. Role of the lever arm in the processive stepping of myosin V. *Proc. Natl Acad. Sci. USA* **99**, 14159–14164 (2002).
- Sakamoto, T. *et al.* Neck length and processivity of myosin V. *J. Biol. Chem.* **278**, 29201–29207 (2003).
- Ruff, C., Furch, M., Brenner, B., Manstein, D. J. & Meyhofer, E. Single-molecule tracking of myosins with genetically engineered amplifier domains. *Nature Struct. Biol.* **8**, 226–229 (2001).
- Rock, R. S. *et al.* Myosin VI is a processive, backwards motor with a large step size. *Proc. Natl Acad. Sci. USA* **98**, 13655–13659 (2001).
- Nishikawa, S. *et al.* Class VI myosin moves processively along actin filaments backward with large steps. *Biochem. Biophys. Res. Commun.* **290**, 311–317 (2002).
- Bahloul, A. *et al.* The unique insert in myosin VI is a structural calcium-calmodulin binding site. *Proc. Natl Acad. Sci. USA* **101**, 4787–4792 (2004).
- Rock, R. S. *et al.* A flexible domain is essential for the large step size and processivity of myosin VI. *Mol. Cell* **17**, 603–609 (2005).
- Homma, K., Yoshimura, M., Saito, J., Ikebe, R. & Ikebe, M. The core of the motor domain determines the direction of myosin movement. *Nature* **412**, 831–834 (2001).
- Coureau, P.-D. *et al.* A structural state of the myosin V motor without bound nucleotide. *Nature* **425**, 419–423 (2003).
- Reubold, T. F., Eschenburg, S., Becker, A., Kull, F. J. & Manstein, D. J. A structural model for actin-induced nucleotide release in myosin. *Nature Struct. Biol.* **10**, 826–830 (2003).
- Coureau, P.-D., Sweeney, H. L. & Houdusse, A. Three myosin V structures delineate essential features of chemo-mechanical transduction. *EMBO J.* **23**, 4527–4537 (2004).
- Holmes, K. C., Schröder, R. R., Sweeney, H. L. & Houdusse, A. The structure of the rigor complex and its implications for the power stroke. *Phil. Trans. R. Soc. Lond. B* **359**, 1819–1828 (2004).
- De La Cruz, E. M., Ostap, E. M. & Sweeney, H. L. Kinetic mechanism and regulation of myosin VI. *J. Biol. Chem.* **276**, 32373–32381 (2001).
- De La Cruz, E. M., Wells, A. L., Rosenfeld, S. S., Ostap, E. M. & Sweeney, H. L. The kinetic mechanism of myosin V. *Proc. Natl Acad. Sci. USA* **96**, 13726–13731 (1999).
- Sweeney, H. L. & Houdusse, A. The motor mechanism of myosin V: insights for muscle contraction. *Phil. Trans. R. Soc. Lond. B* **359**, 1829–1842 (2004).
- Rosenfeld, S. S., Houdusse, A. & Sweeney, H. L. Magnesium regulates ADP dissociation from myosin V. *J. Biol. Chem.* **280**, 6072–6079 (2005).
- Rhoads, A. R. & Friedberg, F. Sequence motifs for calmodulin recognition. *FASEB J.* **11**, 331–340 (1997).
- Meador, W. E., Means, A. R. & Quirocho, F. A. Target enzyme recognition by calmodulin: 2.4 Å structure of a calmodulin–peptide complex. *Science* **257**, 1251–1255 (1992).
- Fallon, J. L. & Quirocho, F. A. A closed compact structure of native Ca²⁺-calmodulin. *Structure* **11**, 1303–1307 (2003).
- Altman, D., Sweeney, H. L. & Spudich, J. A. The mechanism of myosin VI translocation and its load-induced anchoring. *Cell* **116**, 737–749 (2004).
- Tsiavaliaris, G., Fujita-Becker, S. & Manstein, D. J. Molecular engineering of a backwards-moving myosin motor. *Nature* **427**, 558–561 (2004).
- Rice, S. *et al.* A structural change in the kinesin motor protein that drives motility. *Nature* **402**, 778–784 (1999).
- Rosenfeld, S. S., Fordyce, P. M., Jefferson, G. M., King, P. H. & Block, S. M. Stepping and stretching: how kinesin uses internal strain to walk processively. *J. Biol. Chem.* **278**, 18530–18536 (2003).
- Houdusse, A., Kalabokis, V. N., Himmel, D., Szent-Györgyi, A. G. & Cohen, C. Atomic structure of scallop myosin subfragment S1 complexed with MgADP: a novel conformation of the myosin head. *Cell* **97**, 459–470 (1999).
- Yildiz, A. *et al.* Myosin VI steps via a hand-over-hand mechanism with its lever arm undergoing fluctuations when attached to actin. *J. Biol. Chem.* **279**, 37223–37226 (2004).
- Robblee, J. P., Olivares, A. O. & De la Cruz, E. M. Mechanism of nucleotide binding to actomyosin VI: evidence for allosteric head-head communication. *J. Biol. Chem.* **279**, 38608–38617 (2004).
- Sasaki, N., Ohkura, R. & Sutoh, K. Dictyostelium myosin II mutations that uncouple the converter swing and ATP hydrolysis cycle. *Biochemistry* **42**, 90–95 (2003).
- Otwinowski, Z. & Minor, W. Processing X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
- Collaborative Computational Project No. 4., The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
- Navaza, J. AMoRe: an automated package for molecular replacement. *Acta Crystallogr. A* **50**, 157–163 (1994).
- Roussel, A. & Cambillaud, C. *Silicon Graphics Geometry Partner Directory* 77–78 (Silicon Graphics, Mountain View, California, 1989).
- Brünger, A. T. *et al.* Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
- Perrakis, A., Morris, R. M. & Lamzin, V. S. Automated protein model building combined with iterative structure refinement. *Nature Struct. Biol.* **6**, 458–463 (1999).
- Kraulis, P. J. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950 (1991).
- Houdusse, A., Szent-Györgyi, A. & Cohen, C. Three conformational states of scallop myosin subfragment S1. *Proc. Natl Acad. Sci. USA* **97**, 11238–11243 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the staff of the European Synchrotron Radiation Facility for assistance during data collection, A. Li and D. Garbett for technical assistance in preparing the recombinant proteins, and J. Colicari for assistance in crystallization experiments. This work was supported by a grant from the National Institutes of Health (NIAMS) to H.L.S. and A.H., grants from the CNRS and the ARC to A.H., and predoctoral fellowships from the Quebec government and the FRM to A.B.

Author Information Atomic coordinates and structure factors have been deposited in the Protein Data Bank under the accession numbers 2BKH and r2bkhsf for MD^{ins2} and 2BKI and r2bkisf for long MD^{ins2}IQ. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.H. (anne.houdusse@curie.fr) or to H.L.S. (lsweeney@mail.med.upenn.edu).

LETTERS

Release of volatiles from a possible cryovolcano from near-infrared imaging of Titan

C. Sotin¹, R. Jaumann², B. J. Buratti³, R. H. Brown⁴, R. N. Clark⁵, L. A. Soderblom⁶, K. H. Baines³, G. Bellucci⁷, J.-P. Bibring⁸, F. Capaccioni⁹, P. Cerroni⁹, M. Combes¹⁰, A. Coradini⁷, D. P. Cruikshank¹¹, P. Drossart¹⁰, V. Formisano⁷, Y. Langevin⁸, D. L. Matson³, T. B. McCord¹², R. M. Nelson³, P. D. Nicholson¹³, B. Sicardy¹⁰, S. LeMouélic¹, S. Rodriguez¹, K. Stephan² & C. K. Scholz²

Titan is the only satellite in our Solar System with a dense atmosphere. The surface pressure is 1.5 bar (ref. 1) and, similar to the Earth, N₂ is the main component of the atmosphere. Methane is the second most important component², but it is photodissociated on a timescale of 10⁷ years (ref. 3). This short timescale has led to the suggestion that Titan may possess a surface or subsurface reservoir of hydrocarbons^{4,5} to replenish the atmosphere. Here we report near-infrared images of Titan obtained on 26 October 2004 by the Cassini spacecraft. The images show that a widespread methane ocean does not exist; subtle albedo variations instead suggest topographical variations, as would be expected for a more solid (perhaps icy) surface. We also find a circular structure ~30 km in diameter that does not resemble any features seen on other icy satellites. We propose that the structure is a dome formed by upwelling icy plumes that release methane into Titan's atmosphere.

Following the exploration of the jovian system by the Galileo spacecraft, the NASA-ESA-ASI Cassini-Huygens spacecraft went into

orbit in Saturn's system on 1 July 2004. Its mission is devoted to observations of Saturn's atmosphere, rings and satellites. Titan, Saturn's largest satellite, has global characteristics similar to the largest galilean satellites Ganymede and Callisto⁶, which do not have atmospheres. The origin, composition, dynamics and evolution of Titan's atmosphere are key questions to be addressed by the Cassini-Huygens mission.

Scattering by haze particles in Titan's atmosphere makes observation of Titan's surface in the visible very difficult, though it can be easily studied in some narrow infrared windows between the numerous methane absorptions⁷. The Visual and Infrared Mapping Spectrometer (VIMS) instrument, which observes in the 0.35 to 5.2 μm wavelength range, is well positioned to observe Titan's surface⁸. The wavelengths where VIMS has the greatest signal-to-noise ratio while minimizing the effects of scattering by the atmospheric haze are between 2.01 μm and 2.03 μm (Fig. 1), which is one of the seven infrared spectral windows (inset in Fig. 1). The images acquired during this first close fly-by allowed us to build a mosaic showing

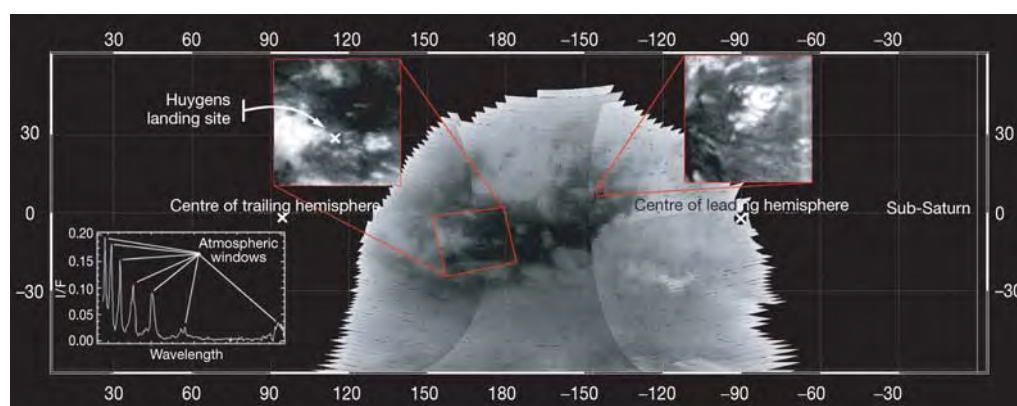


Figure 1 | Titan's map at 2.03 μm wavelength. On 26 October 2004, Cassini-Huygens flew over Titan at less than 1,200 km at closest approach, acquiring several hyperspectral images with spatial resolution ranging from a few tens of kilometres to 2 km per pixel. Titan's diameter is 5,151 km, which is larger than Callisto's diameter (4,806 km) but smaller than Ganymede's diameter (5,268 km). Its mean density is 1,881 kg m⁻³, a value between that of rock and that of water. The VIMS instrument took

hyperspectral images from the visible to 5.2 μm wavelength. This figure shows the mosaic obtained at 2.03 μm ; horizontal and vertical axes show respectively longitude and latitude in degrees. Observations are centred around the anti-Saturn point. Left inset, high-resolution (30 km pixel⁻¹) image, taken to provide the global setting of the site where the Huygens probe successfully landed on 14 January 2005. The best resolved image (right inset) is displayed in Fig. 2.

¹Laboratoire de Planétologie et Géodynamique, UMR CNRS 6112, Université de Nantes, Nantes, 44100, France. ²Institute of Planetary Exploration, DLR, Berlin, 12489, Germany.

³Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109-8099, USA. ⁴Lunar and Planetary Laboratory and Stewart Observatory, University of Arizona, Tucson, Arizona 85721-0092, USA. ⁵US Geological Survey, Denver, Colorado 80225, USA. ⁶US Geological Survey, Flagstaff, Arizona 86011, USA. ⁷Istituto di Fisica dello Spazio Interplanetario, CNR, Rome, 00133, Italy. ⁸Institut d'Astrophysique Spatiale, Université de Paris-Sud, Orsay, 91405, France. ⁹Istituto di Astrofisica Spaziale e Fisica Cosmica, CNR, Rome, 00133, Italy. ¹⁰Observatoire de Paris, Meudon, 92195, France. ¹¹NASA Ames Research Center, Moffett Field, California 94035-1000, USA. ¹²Department of Earth and Space Sciences, University of Washington, Seattle, Washington 98195-1310, USA. ¹³Cornell University, Astronomy Department, Ithaca, New York 14853, USA.

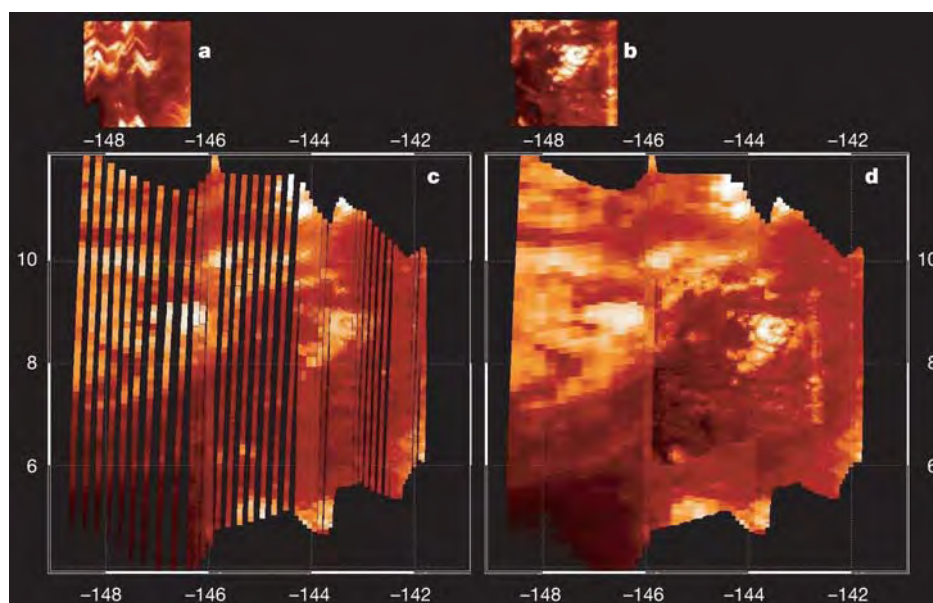


Figure 2 | VIMS high-resolution observations. Two high-resolution images were taken during a 25-min period that ended just before closest approach. Because the spacecraft was moving at 5.8 km s^{-1} (about $20,000 \text{ km h}^{-1}$) relative to Titan, image motion compensation was required so as to point to the same location on Titan's surface. Two images were acquired of the same area 20 min apart. The non-referenced images are presented in **a** and **b**, and the georeferenced images are displayed in **c** and **d**. The first image (**a** and **c**) has a relatively long integration time of 240 ms per pixel. Slight changes in the spacecraft pointing during the observation have distorted the raw image (**a**). Information about Cassini's attitude allowed us to retrieve the exact location of each pixel and to reconstruct the image (**c**). Note that the lines

were not contiguous. The altitude of the spacecraft varied from 11,000 km to about 5,000 km, with a spatial resolution of $5.5 \text{ km pixel}^{-1}$ to $3.0 \text{ km pixel}^{-1}$. The second image was taken with a 80-ms integration, and was centred on the same reference point (**b** and **d**). Because the spacecraft had moved closer to the surface, the resolution varied from $2.6 \text{ km pixel}^{-1}$ to $1.8 \text{ km pixel}^{-1}$. The dominant feature is a bright circular structure (8.5° , -143.5°) with two elongated wings extending westwards. The short-integration-time image (**d**) is put on top of the long-integration image where interpolation has been performed to fill the gaps between the lines. It allowed us to check that the circular feature has not moved between the two shots.

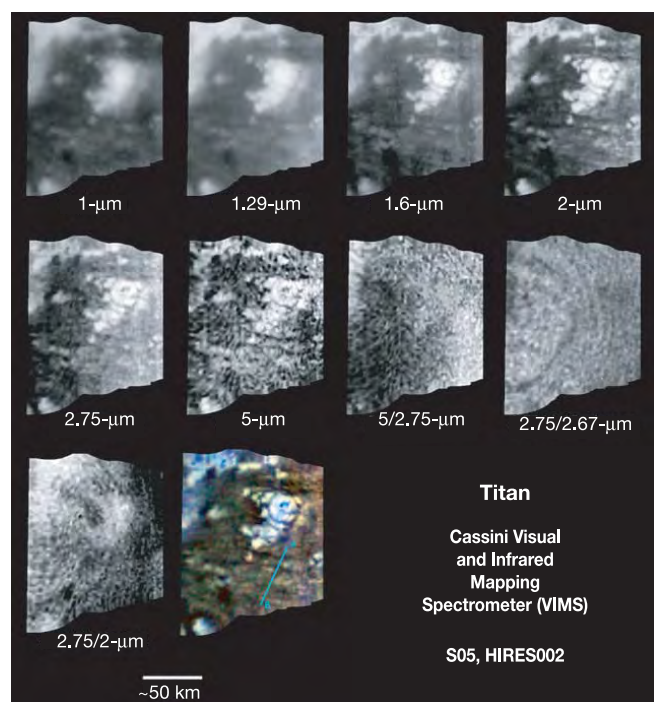


Figure 3 | High-resolution image of Fig. 2d at different wavelengths. Details of the circular feature show up at wavelengths larger than $1.3 \mu\text{m}$. The first six panels are images taken in six infrared windows (inset in Fig. 1), and are georeferenced. Ratio images are represented in order to visualize compositional variations. The last panel is a colour composite image (red, $2.75 \mu\text{m}$; green, $2.0 \mu\text{m}$; blue, $1.6 \mu\text{m}$). On this panel is represented the A–B segment where we perform the topography analysis. This profile crosses east–west lineaments visible in the different images but not in the ratio images.

bright and dark terrains (Fig. 1). A few higher-resolution images, including the Huygens landing site, were obtained as the spacecraft passed closer to Titan.

The highest-resolution image obtained during the 26 October 2004 fly-by of Titan (Fig. 2) covers about $150 \text{ km} \times 150 \text{ km}$. To the right of the image is a bright circular feature about 30 km in diameter, with two elongated wings extending westward. Such a structure resembles volcanic edifices with lobate flows, such as those observed on Earth or Venus for example, although the volcanic material is different. Images in all the infrared windows of the highest-resolution image (Fig. 3) are used to infer information on both the composition and the morphology of this area. At $2.03 \mu\text{m}$ (Fig. 2), the bright regions are about twice as bright as the dark regions, with apparent reflectance (I/F) values of 0.090 and 0.055, respectively. The spectra of both regions look very similar to first order; thus it seems that the bright and dark materials are similar in composition. The window at $2.78 \mu\text{m}$ is wide, with two peaks (inset in Fig. 1). As this spectral region corresponds to the beginning of a strong H_2O absorption band, the presence of water ice should make the ratio of the $2.75 \mu\text{m}$ image to the $2.67 \mu\text{m}$ image (Fig. 3) less than 1, which is not observed, indicating that little, if any, pure water ice is exposed. Also, the $1.9 \mu\text{m}$ window should be present if absorption is only due to methane. The very strong H_2O absorption band at $2 \mu\text{m}$ should make the $2.03 \mu\text{m}$ signal weaker than at $1.9 \mu\text{m}$, which is never the case. These observations suggest that pure H_2O is not dominantly exposed on the surface of the circular feature or the surrounding area.

Information about the morphology can be obtained by comparing images at different wavelengths. If ratio images do not correlate with albedo variations at a given wavelength, the I/F variations at each wavelength are not related to variations in surface composition. One possibility is that these brightness variations are due to illumination of local slopes. In the south of the bright circular feature, the images

at 2 μm , 2.75 μm and 5 μm show east–west linear features that do not show up either in the 5- μm /2.75- μm ratio image, or in the 2.75- μm /2- μm ratio image (Fig. 3). The solar incidence angle is $\sim 34^\circ$ (the lighting is approximately from the lower left). We interpret the many linear features, oriented approximately east–west and typically 20–50 km in length and 5–10 km in separation, to be ridges and valleys (Fig. 4; see legend for details of the method). The right panels in Fig. 4 show the results of the method of estimating the haze-scattering component along profile A–B (shown in the left panel). The last image in Fig. 3 has had the resultant spectral haze model removed. With the haze subtracted, the topographic modulation can be used to make a rough estimate of slopes and, by integration, of the topographic relief across the ridge and valley complex. These ridges and valleys have a relief of the order of a few hundred metres and slopes of $\sim 10^\circ$. If this interpretation is correct, the dark mottled material (Fig. 4) cannot be liquid at a kilometre scale, which is the spatial resolution of this VIMS image.

Because of variations in colour and albedo, the topography in the central part of the bright circular feature cannot be estimated in this way. It is possible that the haze content is not uniform throughout this area (see 1- μm image in Fig. 3). In the centre of the circular feature is a pixel much darker than its surroundings. This pixel is dark in all infrared windows, and suggests a real feature at the surface. Because the illumination comes from the southwest, the difference in brightness can be explained by a depression, thus it is quite tempting to imagine a caldera-like feature in the centre. Circular, darker lineaments also suggest topography.

Finally, we see no obvious evidence for craters on bright or dark areas, meaning that either the surface is very young, or the surface is liquid. The I/F variations seen in the dark mottled material argue strongly against the presence of liquid.

The image of the Huygens landing site (Fig. 1) brings some complementary information. It can first be noted that the bright area has been proven to be at higher elevation than the dark area by the images taken by the Huygens probe during its descent. Within the bright area imaged by the Huygens probe are black lineaments forming a network that has been compared to dry channels. Finally, the landing site, located on the dark area, is not liquid, which confirms the present interpretations.

Several of the features described above suggest that the high-resolution image (Fig. 2) is a zone of extension. The east–west linear features could be interpreted as tectonic features or flow lines. Similar lineaments are found on Ganymede's grooved terrains⁹. Grooved terrains on Ganymede are areas younger than the heavily cratered dark terrains that are thought to result from extension and resurfacing associated with upwelling. If the lineaments on Titan are related to upwelling 'hot ice' and contain contaminants such as hydrocarbons that vaporize as they get closer to the surface, mechanisms similar to those operating for silicate volcanism may produce flows of differentiated ices on Titan that would not be H_2O -based. It is worth noting that tidal heating is an important heat source within Titan, because of this satellite's large eccentricity⁶. If tidal heating is focused in low-viscosity domains (hot ice in the upwelling plumes) as has been hypothesized for Europa¹⁰, this energy may help ice melting and gaseous release in the atmosphere.

The curved black lineaments of the bright circular feature may be similar to dendritic structures seen by Huygens. Such structures could have been formed by the large release of methane-producing rains following the eruptions. If these structures are channels, they would have dried out due to the short timescale for photodissociation of methane in the atmosphere.

Finally, the bright region in the 5- μm /2.75- μm ratio image (Fig. 3) could indicate thermal emission, but the temperature cannot be greater than 200 K. Alternatively, the origin of the bright spot could be related to grain size or compositional variations. Precise radiative-transfer modelling of the atmospheric components and retrieval of

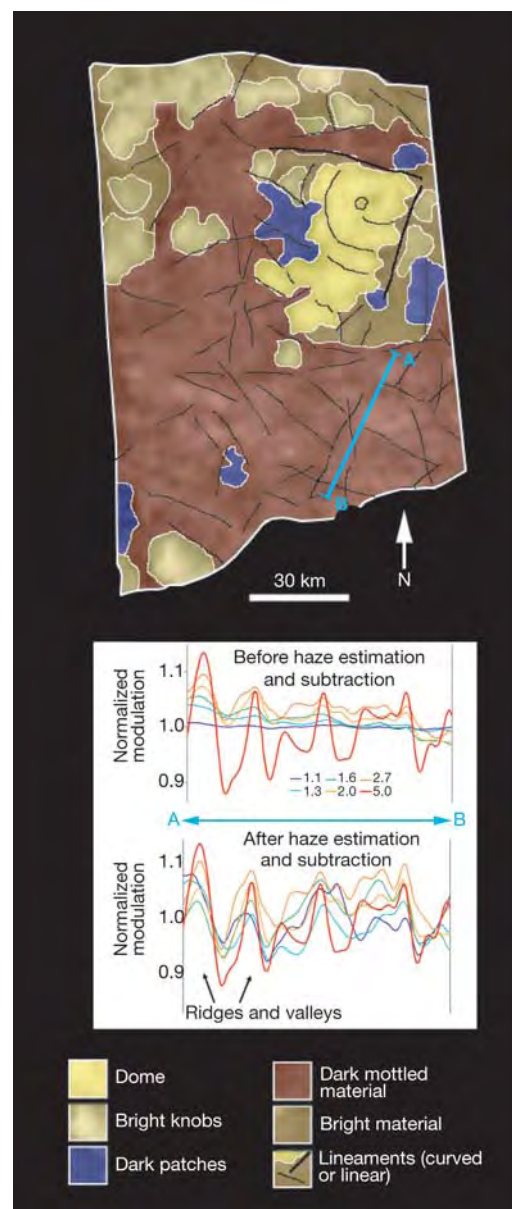


Figure 4 | Geological interpretation of the high-resolution image. On the basis of albedo variations and variations in texture, a geological interpretation of the circular feature has been obtained (left panel). Analysis of I/F variations have been realized along profile A–B. The upper panel and the lower panel on the right shows I/F before and after removal of the haze-scattering effect, respectively. Our interpretation of topography rests on several factors: (1) the highest contrast in the topography is in the cross-sun direction; (2) topographic forms have characteristic bilateral symmetry (bright–dark oscillations in the down-sun direction); and (3) after haze removal at shorter wavelengths, the topographic patterns do not correlate in general with variations in spectral reflectance. In addition to geomorphologic interpretation, we can use the topographic modulation to estimate the scattered radiation in each of the spectral windows. We start with three assumptions: first, the scattering haze has a uniform (additive) contribution over the scene at each wavelength; second, the photometric function is approximately independent of wavelength; and third (most important), the atmospheric scattering at 5.0 μm is zero. Also, a lambertian surface is assumed. Scattering effectively dilutes the topographic modulation. If there were no scattering haze, the topographic modulation (that is, that observed at 5.0 μm) would be the same at all wavelengths. We thereby simply solve for the scattering at each wavelength, which, when subtracted, yields a uniform topographic modulation with wavelength.

surface reflectance are required to distinguish between these different options.

Are there alternative explanations for this bright circular feature? One could imagine that it is a cloud. But there are several arguments against this hypothesis. First, the spectra differ from those already observed for the cloud seen during Cassini's T0 fly-by near Titan's south pole. Second, the two images here (Fig. 2) were taken 20 min apart and no difference can be seen. Third, the Coriolis forces are much smaller on Titan than on Earth owing to Titan's long rotation period (16 days), and circular atmospheric features are not expected at low latitudes.

Another explanation is the accumulation of solid particles transported by gas or liquid, similar to sand dunes on Earth. There is usually a 'V' shape associated with such features, and one could imagine that the bright feature is pointing to the east, with the elongated features representing the tails of the structure. However, the circular shape is difficult to reconcile with such a process. In addition, it would be difficult to explain the dark lineaments with such a structure. Finally, the wind direction in such a case would have to be westward. This is not the direction favoured by current models¹¹ that predict pole to pole circulation, although the direction of winds in a boundary layer of a dense atmosphere can be different.

We interpret this bright circular feature as a cryovolcanic dome in an area dominated by extension. VIMS is an excellent optical instrument for studying the geology of Titan, as below 1.2 μm , atmospheric scattering dominates, and no surface features show up distinctly. Radar observations of the different sites observed by VIMS during this first close fly-by of Titan will be acquired in the course of the mission. They will provide additional information that will help constrain the tectonic setting of the area described in the present study.

Received 27 December 2004; accepted 24 March 2005.

1. Lellouch, E. Atmospheric model of Titan and Triton. *Ann. Geophys.* **8**, 653–660 (1990).
2. Kuiper, G. P. Titan: a satellite with an atmosphere. *Astrophys. J.* **100**, 329–332 (1944).
3. Yung, Y. L., Allen, M. & Pinto, J. P. Photochemistry of the atmosphere of Titan: comparison between model and observations. *Astrophys. J. Suppl.* **55**, 465–506 (1984).
4. Lunine, J. I. Does Titan have an ocean? A review of current understanding of Titan's surface. *Rev. Geophys.* **31**, 133–149 (1993).
5. Lunine, J. I., Stevenson, D. J. & Yung, Y. L. Ethane ocean on Titan. *Science* **222**, 1229–1230 (1983).
6. Sotin, C. & Tobie, G. Internal structure and dynamics of the large icy satellites. *C.R. Acad. Sci.* **5**, 769–780 (2004).
7. Meier, R., Smith, B. A., Owen, T. C. & Terrile, R. J. The surface of Titan from NICMOS observations with the Hubble Space Telescope. *Icarus* **145**, 462–473 (2000).
8. Brown, B. *et al.* Observations with the Visual and Infrared Mapping Spectrometer (VIMS) during Cassini's flyby of Jupiter. *Icarus* **164**, 461–470 (2003).
9. Lucchita, B. K. Grooved terrains on Ganymede. *Icarus* **44**, 481–501 (1980).
10. Sotin, C., Head, J. W. & Tobie, G. Europa: Tidal heating of upwelling thermal plumes and the origin of lenticulae and chaos melting. *Geophys. Res. Lett.* **29**(8), doi:10.1029/2001GL013844 (2002).
11. Gibbard, S. G. *et al.* Titan's 2- μm surface albedo and haze optical depth in 1996–2004. *Geophys. Res. Lett.* **31**, doi:10.1029/2004GL019803 (2004).

Acknowledgements We thank E. Mercier, D. Mège, J.-P. Combe and O. Bourgeois for discussions about interpreting the high-resolution image, and R. Wagner for help in the projection of the cubes.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.S. (Christophe.Sotin@univ-nantes.fr).

LETTERS

Discovery of an aurora on Mars

Jean-Loup Bertaux¹, François Leblanc¹, Olivier Witasse², Eric Quemerais¹, Jean Lilensten³, S. A. Stern⁴, B. Sandel⁵ & Oleg Korablev⁶

In the high-latitude regions of Earth, aurorae are the often-spectacular visual manifestation of the interaction between electrically charged particles (electrons, protons or ions) with the neutral upper atmosphere, as they precipitate along magnetic field lines. More generally, auroral emissions in planetary atmospheres “are those that result from the impact of particles other than photoelectrons” (ref. 1). Auroral activity has been found on all four giant planets possessing a magnetic field (Jupiter², Saturn³, Uranus⁴ and Neptune⁵), as well as on Venus, which has no magnetic field⁶. On the nightside of Venus, atomic O emissions at 130.4 nm and 135.6 nm appear in bright patches of varying sizes and intensities⁶, which are believed to be produced by electrons with energy <300 eV (ref. 7). Here we report the discovery of an aurora in the martian atmosphere, using the ultraviolet spectrometer SPICAM on board Mars Express. It corresponds to a distinct type of aurora not seen before in the Solar System: it is unlike aurorae at Earth and the giant planets, which lie at the foot of the intrinsic magnetic field lines near the magnetic poles, and unlike venusian auroras, which are diffuse, sometimes spreading over the entire disk. Instead, the martian aurora is a highly concentrated and localized emission controlled by magnetic field anomalies in the martian crust.

Mars has no intrinsic internal magnetic field, so no aurora of the Earth/Saturn/Jupiter type is expected. However, it was suggested some years ago that a Venus-like aurora could exist⁸. More recently, since the discovery by Mars Global Surveyor of Mars’ crustal magnetic field⁹, a remnant of an ancient intrinsic field, it was hypothesized that particles could precipitate along those magnetic field lines connecting the surface to the solar wind^{10–12}. We describe here the observation of such an aurora produced by precipitating particles, which has been made by the SPICAM instrument on board the ESA Mars Express mission. SPICAM (Spectroscopy for the Investigation of the Characteristics of the Atmosphere of Mars) consists of separate ultraviolet (UV) and near-infrared spectrometers dedicated primarily to the study of the atmosphere of Mars^{13,14}. SPICAM operates in various modes: nadir viewing, limb viewing and atmospheric vertical profiling by stellar and solar occultations¹⁵. It measures simultaneously in five adjacent fields of view (FOV or spatial bins), two of them with a higher spectral resolution and lower sensitivity (spatial bins 1 and 2). In the observation reported here, the SPICAM line of sight (LOS) was permitted to drift slowly across the nightside limb to search for any weak emission, as no martian nightglow had been reported before. (For further details of the SPICAM instrument and observing geometry, see Methods.)

Figure 1a is a colour-coded image of the time series of high-resolution intensity spectra. The most obvious spectral features are the H Lyman- α emission at 121.6 nm, and a well-structured band in the region 190–270 nm, of variable intensity but with shapes that appear constant. This emission was discovered by SPICAM and

identified recently¹⁶ as the γ and δ bands of the nitric oxide (NO) molecule, emitted when O and N atoms recombine, after having been produced by solar extreme UV photo-dissociation of CO₂, O₂ and N₂ on the dayside and transported to the nightside.

In Fig. 1b the nightglow signal integrated over the wavelength range of the NO bands (181–298 nm) is displayed as a function of time for the five spatial bins. The signal is more intense for spatial bins 3, 4 and 5 than for spatial bins 1 and 2 because the FOV is wider and the source is extended. All curves show the same behaviour, almost identical to the variation of the NO emission observed six days later at orbit 734¹⁶; this behaviour is explained by the variation of the altitude and the latitude of the Mars nearest point (MNP) when the LOS scans across the NO emitting layer—this layer is confined in the altitude range 60–80 km, and is more intense at large southern latitudes (around time 750 s). There is however also a strong peak in all spatial bins between times 533 and 540 s (an increase by a factor of 3 to 4), which has no equivalent during orbit 734. The time series of spectra (Fig. 1a) shows that the spectrum of this extra source of light differs from the typical NO spectrum. This emission was observed when the MNP of the FOV was at a tangent altitude of 19 km, where the local time was 21 h, and its longitude and latitude were 198.4° and –46.3°, respectively.

The detailed peak signals are displayed in Fig. 2 for the five spatial bins aligned horizontally. The spacecraft horizontal velocity v was 4.3 km s^{–1}, perpendicular to the LOS. There is a time delay of 2.3 s between spatial bin 5 and spatial bin 1 signals (better seen at exit than at entry), while the (horizontal) angular size of each FOV is 0.32° (5.5 mrad). This corresponds to an apparent angular velocity $\omega = 9.56$ mrad s^{–1}. Therefore, the distance d to the light emitting volume (an aurora, as we shall see) is found to be $d = v/\omega = 450$ km (assuming the aurora is unchanging during these few seconds).

The horizontal size of the aurora can be inferred from the time it is seen by each spatial bin: about 7 s (best determined for bins 3, 4 and 5), giving a size of 30 km for the total size of the emission region (taking into account the finite size of one FOV bin). The uncertainty on this time is estimated to be less than 0.5 s, which implies an uncertainty on d of about 50 km and an uncertainty on the size of the emitting region of about 2 km. Considering that the altitude of the spacecraft at the moment of this observation was 266 km above the surface, and that MNP altitude was 19 km at 1,325 km from the spacecraft, geometry implies that the altitude of the observed emission is 129 ± 13 km. Still, it could extend further vertically, as it was scanned horizontally.

The two spectra shown in Fig. 3 are photometrically calibrated by comparison with hot stars emitting in the UV¹³. The spectrum in Fig. 3a displays the already reported feature of the martian nightglow—that is, the NO γ and δ bands, and the H Lyman- α emission. The main emissions in the net auroral spectrum (Fig. 3b) are the CO $a^3\Pi-X^1\Sigma$ Cameron band between 180 and 240 nm, long observed on

¹Service d’Aéronomie du CNRS/IPSL, BP 3, Verrières-le-Buisson, 91371, France. ²ESA-ESTEC, Postbus 299, NL 2200AG, Noordwijk, The Netherlands. ³Laboratoire de Planétologie de Grenoble, CNRS, Université J. Fourier, B.P. 53, 38041, Grenoble, France. ⁴Southwest Research Institute, 1050 Walnut Street, Boulder, Colorado 80302, USA. ⁵Lunar and Planetary Laboratory, 1541 E. University Boulevard, University of Arizona, Tucson, Arizona 85721, USA. ⁶Space Research Institute (IKI), 84/32 Profsoyuznaya, 117810 Moscow, Russia.

the martian dayside¹⁷, as well as emissions associated with atomic carbon resonances and with the CO $A^1\Pi-X^1\Sigma^+$ fourth positive group between 135 and 170 nm, emissions associated with the $\text{CO}_2^+ B^2\Sigma_u^+ - X^2\Pi_g$ doublet at 289 nm and with O at 297.2 nm. Lyman- α emission has been suppressed in this figure by subtracting the spectrum of Fig. 3a (which also indicates that no increase of this emission is observed during this time). O emissions at 130.4 and 135.6 nm, if present, are at or below the limit of the sensitivity of the instrument, and would need further observations to be confirmed. The peak at 276 nm (54 ± 38 R, where R indicates Rayleighs) has not been reported in Mariner 6, 7 and 9 observations. It could correspond to N_2 Vegard Kaplan (0,6) band emission¹⁸ or to O_2 Herzberg I (6,2) band emission¹⁹. O_2 Herzberg I emission is, however, less likely, as it should have been also seen in Fig. 3a, as a result of O + O

recombination, similarly to NO emission. The intensities of the strongest emission are for the Cameron band, 694 ± 50 R; for $\text{CO}_2^+ B^2\Sigma_u^+ - X^2\Pi_g$, 71 ± 42 R; and for O at 297.2 nm, 23 ± 46 R. The presence of such emissions in the night therefore indicates excitation of the martian atmosphere by a flux of particles, probably electrons (the possibility that these emissions are produced by protons or other charged particles cannot be firmly excluded).

Figure 4 displays the variation of the geographic position of the spacecraft during this orbit (white dashed line) and also the position of the MNP of the LOS (solid white line), superimposed on a map of the radial component of the crustal magnetic field B as reported by Mars Global Surveyor⁹. It is notable that the region of strongest crustal magnetic field B is at about the location along the LOS determined above for the auroral emission. At the time of the

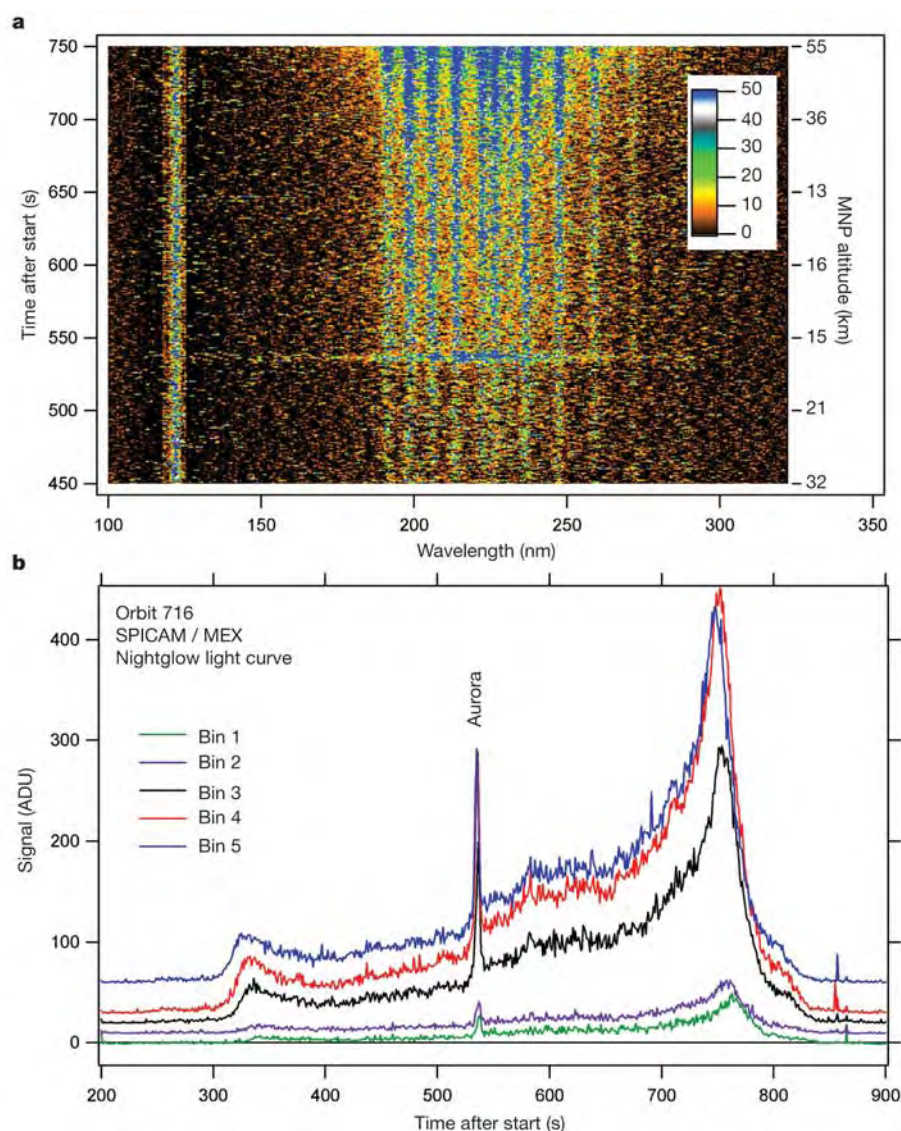


Figure 1 | Time variation of the martian nightglow intensity. **a**, Spectra recorded during the grazing limb observation with spatial bin 2 (narrow slit, spectral resolution ~ 1.5 nm, one spectrum per second, data subset from 450 to 750 s after the start of observation). It contains the H Lyman- α emission at 121.6 nm and a well structured band (190–270 nm), identified¹⁶ as NO γ and δ bands (Fig. 3). The intensity in ADU (analogue to digital units) per pixel is colour-coded. CCD dark current has been subtracted. Altitudes of the Mars nearest point (MNP) of the line of sight are indicated at the right. At the time of the peak marked 'Aurora' in **b**, the spectra are obviously different from the typical NO spectrum. **b**, Signal intensity for all five spatial

bins as a function of time between 200 and 900 s after the start of observations. Units are ADU per spectral pixel = 0.54 nm, averaged from 181 to 298 nm. There are 37 ADU per detected photon for the particular high voltage used here. The curves for spatial bins 2, 3, 4 and 5 have been vertically displaced for clarity (respectively by 10, 20, 30 and 60 ADU). Spatial bins 3, 4 and 5 have low resolution but high sensitivity, and bins 1 and 2 are less sensitive but have higher spectral resolution. A conspicuous spike marked 'Aurora' is observed in all bins at time 535 s. This is the time at which the spectra in **a** differ from the usual NO spectrum.

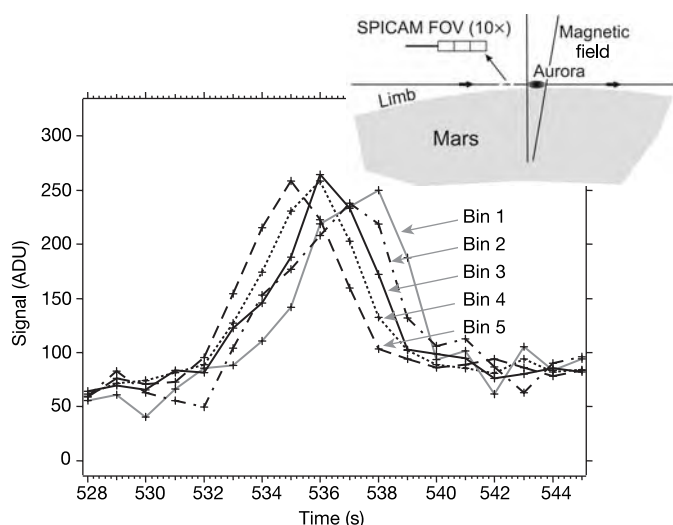


Figure 2 | Detail of the time variation of the signal for the various spatial bins. They are computed after normalization to the most intense spatial bin by their relative sensitivities, obtained by integration of the whole light curves. There is a delay in the light curves corresponding to the sketch (inset): the aurora is first seen in the field of view (FOV) of bin 5, then bin 4, and so on. It leaves spatial bin 5 first, and so on. From the angular size of the FOV, and the known orbital velocity of the spacecraft (perpendicular to the LOS), the horizontal size of the aurora is determined to be 30 ± 2 km. The vertical extent of the aurora remains unknown.

brightest signal reported above, the line of sight was passing through the largest martian magnetic crustal anomalies (dark red region in the middle-left of Fig. 4). These anomalies were at a distance of between 300 and 600 km from the satellite, in excellent agreement with the distance to the source inferred from the timing of the observation. In the measurements by the electron reflectometer on board Mars Global Surveyor during the mapping orbit (400 km in altitude)¹¹, electron spikes were often observed on the martian nightside near a local maximum in the radial component of the crustal field. Such events correspond to a sharp increase of the flux of electrons. The duration of the spikes was usually more than 8 s, during which the spacecraft moved at least 27 km along its orbit. The

close correlation between the location of our observed emission and the position of the crustal field anomalies indicates that the origin of these emissions is a flux of electrons moving along the crustal magnetic field lines and exciting the upper atmosphere of Mars. Using the crustal magnetic field measurements by Mars Global Surveyor⁹, Krymskii *et al.*¹² concluded that the region where SPICAM observed the aurora displayed in Fig. 3b (177° longitude, -52° latitude) should indeed correspond to a cusp type region (their figure 8).

To estimate the intensity of the flux of electrons needed to produce the observed intensity, we use TRANSCAR model^{20–22}, which describes the energy deposition of electrons through the martian

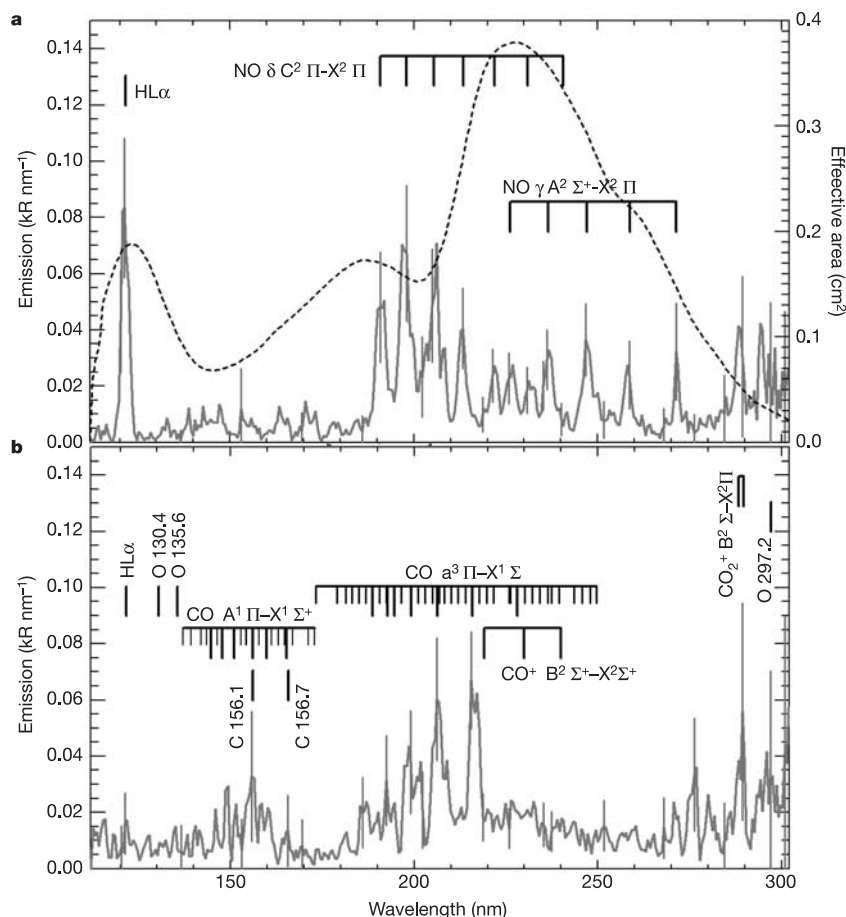


Figure 3 | Two different types of spectra. **a**, Spectrum formed by averaging spectra obtained just after the aurora between times of observation 541–548 s after the start of the observation, typical of the NO emission γ and δ bands, and H Lyman- α emission. The dashed curve is the sensitivity of SPICAM as a function of wavelength, expressed as the effective area (in cm^2 , right scale). It is defined as the ratio of the number of detected photons (photo-events) to the fluence of photons entering the instrument (photons cm^{-2}). **b**, Net auroral emission. Spectrum obtained by averaging spectra between times of observation 533–540 s after the start of the observation, from which was subtracted spectrum **a**. Only data from the narrow slit have been used in this calculation (best spectral resolution). A number of known spectroscopic transitions are indicated. The CO Cameron band is clearly seen. Other emissions are discussed in the text. Intensity error bars (1σ) are indicated for a few spectral pixels as thin vertical segments.

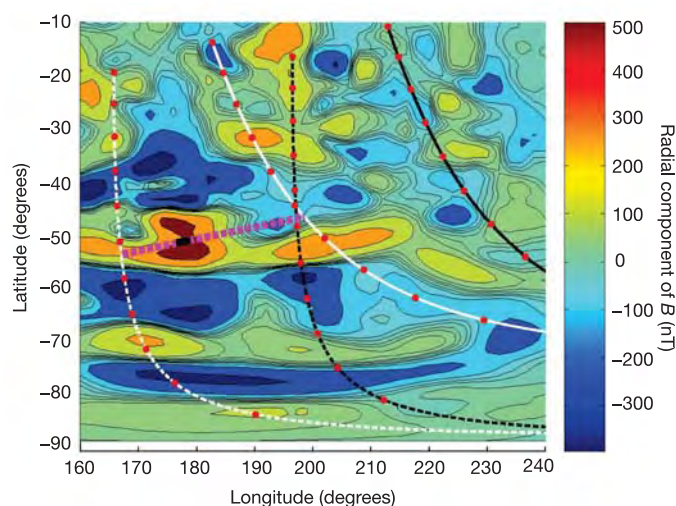


Figure 4 | Map of the radial component of B , the martian crustal magnetic field, at 200 km altitude. Values are deduced from Mars Global Surveyor aerobraking and science phasing orbits²⁵. White dashed line, Mars Express position during orbit 716 (starts at the top); white solid line, position of the MNP. Black dashed and solid lines show Mars Express position and MNP for orbit 734, during which no auroral emission was recorded. Magenta dashed line, LOS of SPICAM at 535 s after start; the dark rectangle on this line is our estimate of the position of the centre of the aurora along the LOS, and its uncertainty. The auroral region might be more elongated along this LOS.

atmosphere associated with ionospheric transport, and deduces from this calculation some of the martian emissions (in particular, $\text{CO}_2^+ \text{B}^2\Sigma_u^+ - \text{X}^2\Pi_g$ at 289.7 nm, which is well defined in Fig. 3b). The neutral atmosphere is taken to be similar to the dayside case at solar minimum conditions (from Viking observations²³), and the magnetic field is assumed to be fully radial.

The energy distributions measured by the electron reflectometer of Mars Global Surveyor are set to be the upper limit of the domain of calculation. Using the typical magnetosheath spectrum measured by the electron reflectometer on board Mars Global Surveyor¹¹, the corresponding $\text{CO}_2^+ \text{B}^2\Sigma_u^+ - \text{X}^2\Pi_g$ emission at 289.7 nm for limb scanning with a path length of 30 km reaches a maximum at 167 km and varies between 0.01 and 4.2 R from 117 to 142 km (with 3 R at 130 km). For the typical magnetotail electron distribution¹¹, such an emission is reduced to a maximum of 0.3 R at 140 km and decreases to 0.05 R at 117 km.

When comparing the observed emission to a model prediction of the emissivity ($\text{photons cm}^{-3} \text{s}^{-1}$), the main unknown is the length along the LOS of the emission zone. In the case of the typical magnetosheath electron flux, a flux of electrons one order of magnitude more intense than the typical flux reported by Mars Global Surveyor is needed to produce the observed emission intensity, if the size of the aurora is 30 km along the LOS, or it could be produced by the reported flux, if the aurora is at least 300 km in horizontal extent. In the case of a magnetotail electron flux, which peaks at about the same altitude observed by SPICAM for the auroral feature, a flux of electrons two orders of magnitude more intense than the typical flux reported by Mars Global Surveyor is needed (a 3,000 km emitting region in such a case is not compatible with the size of the crustal field anomalies). Haider *et al.*²⁴ calculated the emission expected to arise from electrons penetrating the martian atmosphere on the nightside. Using a typical energy distribution for the magnetotail electrons as reported by the Phobos mission (which is actually in better agreement with the magnetosheath electron distribution from Mars Global Surveyor¹¹), these authors estimated CO Cameron band emission of ~ 60 R at the nadir. Assuming that the vertical extent of the aurora is of the order of a few scale heights (~ 8 km), viewing the aurora at the nadir would yield an intensity

similar to that of viewing it horizontally with a path length of 30 km. Haider *et al.*²⁴ also calculated a nadir intensity of 30 R for the green atomic oxygen line at 557.7 nm. From this value can be derived an O I 297.2 nm intensity roughly 15 times smaller than the 557.7 nm intensity, that is, 2 R, in very good agreement with the 2.1 R estimated by TRANSCAR. Therefore the calculation by Haider *et al.*²⁴ of the expected CO Cameron band emission can also be used to infer that the emission observed by SPICAM is most likely to be due to a flux of precipitating electrons with an intensity one order of magnitude larger than the typical magnetosheath flux measured by Mars Global Surveyor, or to an emitting region at least 30×300 km in extent, elongated east–west.

Mars Global Surveyor measured the electron flux at 400 km altitude, and the emission reported here is produced below 140 km altitude. Mitchell *et al.*¹¹ reported that the flux spikes persisted for at least 8 s, that is, they were observed over a distance larger than 27 km. In the present case, the magnetic flux tube at the origin of this emission is most probably associated with the strongest crustal anomalies reported by Mars Global Surveyor⁹. According to the SPICAM observation, the horizontal size of the magnetic flux tube at 140 km should be about 30 ± 2 km. One possible way to induce a one order of magnitude more intense flux at 140 km, given the flux measured at 400 km, is for the cross-section of the magnetic flux tube to decrease by a factor of 10 from 400 to 140 km altitude. Then the flux tube at 400 km would have a diameter of 100 km, which is in rather good agreement with the size of the cusp region as constructed by Krymskii *et al.*¹².

The Mars Express SPICAM observation reported here provides for the first time key insight into the role of the crustal magnetic field in producing real cusp-like structures at Mars. Such a cusp structure concentrates the incident electron flux into a small region of the martian atmosphere, induces local auroral emission and should also induce a local maximum in ionospheric density. The absence of any detection of auroral emission at other places during orbits 716 and 734 (see Fig. 4) indicates that a much smaller electron flux reaches most of the martian atmosphere on the nightside.

METHODS

Description of the instrument. SPICAM's UV channel is a full UV imaging spectrometer (118–320 nm, spectral resolution 1.5 nm, intensified CCD). The telescope is a single off-axis parabolic mirror, at whose focus is a slit, which acts both to define the FOV and is the entrance slit of the spectrometer. This spectrometer uses a holographic toroidal grating, with the dispersion direction perpendicular to the slit. Therefore, each point of the slit has its spectrum formed perpendicular to the slit, on the photocathode of an image intensifier (Hamamatsu) with a CsTe cathode (UV sensitive, but blind to visible light), microchannel plate electron multiplier, and a phosphor output screen. The green image from the phosphor is transferred by fibre optic coupling to a Thompson CCD (384×288 pixels). The 288 lines of the CCD are oriented along the spectral direction, and each line records the spectrum of one point of the entrance slit, with 384 spectral elements. The focal length of the telescope is such that one CCD pixel covers an FOV of $0.01 \times 0.01^\circ$.

The slit of the spectrometer has two parts: a narrow part (0.02° wide by 1.9° long), achieving a spectral resolution of 1.5 nm (about 3 CCD pixels), and a wide part (0.2° wide by 0.98° long), achieving a higher photometric sensitivity for extended sources (by a factor of ~ 8), at the expense of a reduced spectral resolution (6 nm). In principle, SPICAM can record 288 spatially resolved spectra along its 2.88-degree-long slit (that is, each spatial element subtending 0.02×0.01 degrees on the sky). However, in practice, only five spectra are transmitted for each 1 s measurement in order to save data volume transmission. These are usually a sum of N individual spectra, with $N = 32$ in the present case, forming five adjacent spatial bins (numbered from 1 to 5, each extending along 0.32° on the sky, and in total 1.6° along the slit direction). Two such spatial bins 1 and 2 were acquired along the narrow slit and the three others (3, 4 and 5) with the wide slit, with total solid angle of 0.02×0.32 and 0.2×0.32 square degrees, respectively.

Geometry of limb grazing observations. At the time of the auroral observation (orbit 716, 11 August 2004, start time 05:52:30 UT), the orbit of Mars Express had its pericentre on the nightside (18:55 local time, at a solar zenith angle of 117.5°), at an altitude of 266 km, and latitude -55° . The orbit is almost polar, and Mars

Express was descending from north to south (Fig. 4). The spacecraft orientation was fixed in an inertial system so that the LOS of SPICAM (the middle of spatial bin 3) was perpendicular to the velocity vector at pericentre, looking sideways. Therefore, it provided a grazing view of the limb while Mars Express was passing through its pericentre (± 10 min). As a result, the altitude of the tangent point of the LOS (the MNP) decreased from 371 km (05:52:29 UT), to a minimum of 17.8 km at pericentre (06:01:54 UT) and then increased back to 428 km (06:12:09 UT). At the same time, the tangent point latitude changed from -13° to -69° .

The whole altitude range at the limb was therefore scanned twice, but at different latitudes on the descending and ascending branches. The spectrometer slit was maintained parallel to the limb as seen from pericentre, that is, it was horizontal (Fig. 2). Projected at the limb distance of $\sim 1,300$ km, the instantaneous FOVs of the spatial bins correspond to 0.46×7.4 km and 4.6×7.4 km respectively for the narrow (1 and 2) and wide slit spatial bins (3, 4, 5). However, as discussed in the text, the aurora is estimated to be at a distance of 450 km from Mars Express, where the FOVs cover only 0.16×2.5 km and 1.6×2.5 km, respectively.

Received 3 February; accepted 29 March 2005.

1. Fox, J. L. Models for aurora and airglow emissions from other planetary atmospheres. *Can. J. Phys.* **64**, 1631–1656 (1986).
2. Broadfoot, A. L. *et al.* Extreme ultraviolet observations from Voyager 1 encounter with Jupiter. *Science* **204**, 979–982 (1979).
3. Broadfoot, A. L. *et al.* Extreme ultraviolet observations from Voyager 1 encounter with Saturn. *Science* **212**, 206–211 (1981).
4. Bhardwaj, A. & Gladstone, G. R. Auroral emissions of the giant planets. *Rev. Geophys.* **38**, 295–353 (2000).
5. Broadfoot, A. L. *et al.* Ultraviolet spectrometer observations of Neptune and Triton. *Science* **246**, 1459–1466 (1989).
6. Philips, J. L., Stewart, A. I. F. & Luhmann, J. G. The Venus ultraviolet aurora: observations at 130.4 nm. *Geophys. Res. Lett.* **13**, 1047–1050 (1986).
7. Fox, J. L. & Stewart, A. I. F. The Venus ultraviolet aurora: A soft electron source. *J. Geophys. Res.* **96**, 9821–9828 (1991).
8. Fox, J. L. in *Venus and Mars: Atmospheres, Ionospheres, and Solar Wind Interactions* (eds Luhmann, J. G., Tatrallyay, M. & Pepin, R. O.) 191–222 (Geophys. Monogr. 66, American Geophysical Union, Washington DC, 1992).
9. Acuña, M. H. *et al.* Magnetic field of Mars: Summary of results from the aerobraking and mapping orbits. *J. Geophys. Res.* **106**, 23403–23417 (2001).
10. Ness, N. F. *et al.* Effects of magnetic anomalies discovered at Mars on the structure of the Martian ionosphere and solar wind interaction as follows from radio occultation experiments. *J. Geophys. Res.* **105**, 15991–16004 (2000).
11. Mitchell, D. L. *et al.* Probing Mars' crustal magnetic field and ionosphere with the MGS Electron Reflectometer. *J. Geophys. Res.* **106**, 23419–23427 (2001).
12. Krymskii, A. M. *et al.* Structure of the magnetic field fluxes connected with crustal magnetization and topside ionosphere at Mars. *J. Geophys. Res.* **107**, 1245, doi:10.1029/2001JA000239 (2002).
13. Bertaux, J. L. *et al.* The study of the Martian atmosphere from top to bottom with SPICAM Light on Mars Express. *Planet. Space Sci.* **48**, 1303–1320 (2000).
14. Bertaux, J. L. *et al.* in *Mars Express* (ed. Chicarro, A.) 95–120 (Special Publication SP1240, ESA, Noordwijk, 2004).
15. Bertaux, J. L. *et al.* Global structure and composition of the martian atmosphere with SPICAM on Mars Express. *Adv. Space Res.* **35**, 31–36 (2005).
16. Bertaux, J.-L. *et al.* First observation of nightglow in the upper atmosphere of Mars: the NO bands in UV and implications for atmospheric transport. *Science* **307**, 566–569 (2005).
17. Barth, C. A. *et al.* Mariner 6 and 7 ultraviolet spectrometer experiment: Upper atmosphere data. *J. Geophys. Res.* **76**, 2213–2227 (1971).
18. Fox, J. L. & Dalgarno, A. Ionization, luminosity, and heating of the upper atmosphere of Mars. *J. Geophys. Res.* **84**, 7315–7333 (1979).
19. Eastes, R. W., Huffman, R. E. & Leblanc, F. J. NO and O₂ ultraviolet nightglow and spacecraft glow from the S3–4 satellite. *Planet. Space Sci.* **40**, 481–493 (1992).
20. Blelly, P.-L., Witasse, O. & Lilensten, J. A coupled kinetic, fluid and MHD model of the dayside ionosphere of Mars including the energetics. I: General description. *J. Geophys. Res.* (submitted).
21. Lilensten, J., Witasse, O., Blelly, P.-L., Bougher, S. W. & Engel, S. A coupled kinetic, fluid and MHD model of the dayside ionosphere of Mars including the energetics. II: Study of the ion production and of the CO₂⁺ (B-X) emission band. *J. Geophys. Res.* (in the press).
22. Witasse, O., Blelly, P.-L., Lilensten, J., Engel, S. & Bougher, S. W. A coupled kinetic, fluid and MHD model of the dayside ionosphere of Mars including the energetics. III: Viking 1 lander and Mariner 6 cases studies. *J. Geophys. Res.* (in the press).
23. Nier, A. O. & McElroy, M. B. Composition and structure of Mars' upper atmosphere — results from the neutral mass spectrometers on Viking 1 and 2. *J. Geophys. Res.* **82**, 4341–4349 (1977).
24. Haider, S. A. *et al.* Calculated ionization rates, ion densities, and airglow emission rates due to precipitating electrons in the nightside ionosphere of Mars. *J. Geophys. Res.* **97**, 10637–10641 (1992).
25. Purucker, M. *et al.* An altitude-normalized magnetic map of Mars and its interpretation. *Geophys. Res. Lett.* **27**, 2449–2452 (2000).

Acknowledgements Mars Express is a space mission from ESA (European Space Agency). We thank all ESA members who participated in this mission, in particular M. Denis at ESOC for the controlling of the spacecraft, and R. Pischel and T. Zeghers at ESTEC for planning exercises; Astrium Corp. for the design and construction of the spacecraft; our collaborators at the three institutes for the design and fabrication of the instrument (Service d'Aéronomie/France, BIRA/Belgium and IKI/Moscow), and in particular E. Dimarellis as SPICAM Project Manager and E. Van Ransbeeck at BIRA for mechanical design and fabrication. We also thank CNRS and CNES for financing SPICAM in France, the Belgian government, the Russian Academy of Sciences and NASA for support of US co-investigators. B.R.S. was supported by the Jet Propulsion Laboratory.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.-L.B. (bertaux@aerov.jussieu.fr).

A thermally self-sustained micro solid-oxide fuel-cell stack with high power density

Zongping Shao¹, Sossina M. Haile¹, Jeongmin Ahn², Paul D. Ronney², Zhongliang Zhan³ & Scott A. Barnett³

High energy efficiency and energy density, together with rapid refuelling capability, render fuel cells highly attractive for portable power generation^{1,2}. Accordingly, polymer-electrolyte direct-methanol fuel cells are of increasing interest as possible alternatives to Li ion batteries³. However, such fuel cells face several design challenges and cannot operate with hydrocarbon fuels of higher energy density. Solid-oxide fuel cells (SOFCs) enable direct use of higher hydrocarbons^{4–6}, but have not been seriously considered for portable applications because of thermal management difficulties at small scales, slow start-up and poor thermal cyclability. Here we demonstrate a thermally self-sustaining micro-SOFC stack with high power output and rapid start-up by using single chamber operation on propane fuel. The catalytic oxidation reactions supply sufficient thermal energy to maintain the fuel cells at 500–600 °C. A power output of ~350 mW (at 1.0 V) was obtained from a device with a total cathode area of only 1.42 cm².

Single-chamber fuel cells (SCFCs) are a type of fuel cell, which utilize a fuel + O₂ mixture gas (diluted) as the same gas atmosphere

for both the anode and cathode^{7–11}. The performance of single-chamber SOFCs is based on the highly selective anode and cathode towards the gas mixture. In addition to its role as the electrochemical catalyst for H₂ and CO oxidation, the anode also acts as the non-electrochemical catalyst for partial oxidation of the hydrocarbon fuel to form these more electrochemically active species. The oxidation reaction is exothermic in nature (for example, C₃H₈(g) + 3/2O₂(g) = 3CO(g) + 4H₂(g), $\Delta H_{550\text{ }^{\circ}\text{C}} = -206.96\text{ kJ mol}^{-1}$); the heat released raises the temperature of the fuel cells beyond that of the furnace in which they are placed^{12,13}. We used this heat to sustain the fuel-cell temperature in the absence of external heating. Of the fuel choices available for portable power applications, propane is particularly attractive because, as a gas at temperatures down to -42 °C at ambient pressure, it can be easily supplied to a power-generation device, and as liquid phase under moderate pressure (~10 atm at 25 °C), it can readily be stored with high energy density (6.3 kW h l⁻¹ at 25 °C, about twice that of methanol).

Anode-supported, thin-film electrolyte fuel cells were fabricated as reported elsewhere using samaria-doped ceria (Sm_{0.15}Ce_{0.85}O_{1.925},

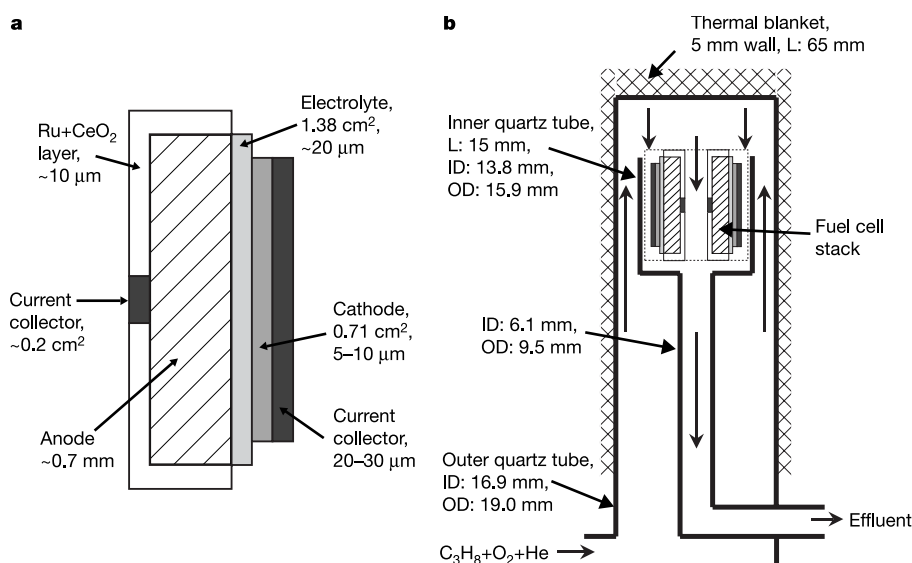


Figure 1 | Schematic diagram of the cell configuration and reactor system used in the thermally self-sustaining micro-SOFCs. **a**, The disc-shaped fuel cell comprises an SDC electrolyte (~20 μm) supported on an Ni + SDC anode (0.7 mm) with BSCF + SDC (70:30 wt%) as the cathode (5–10 μm); a ~10-μm-thick Ru + CeO₂ catalyst layer was coated onto the entire anode surface, with the exception of about 0.2 cm² at the centre, over which the

current collector (Ag paste) was placed; the entire cathode surface was covered with a layer of BSCF + Ag which served as the current collector (20–30 μm). **b**, The fuel-cell reactor comprises a removable ceramic thermal insulator or blanket, an inner quartz tube, and an outer quartz tube for introducing the gas mixture for the fuel cell. Component dimensions (L, width; ID, inner diameter; OD, outer diameter) are as indicated.

¹Materials Science, California Institute of Technology, Pasadena, California 91125, USA. ²Department of Aerospace and Mechanical Engineering, University of Southern California, Los Angeles, California 90089, USA. ³Department of Materials Science and Engineering, Northwestern University, Evanston, Illinois 60208, USA.

SDC) as the electrolyte, Ni + SDC as the anode (SDC:NiO = 40:60 by weight), and a mixture of $\text{Ba}_{0.5}\text{Sr}_{0.5}\text{Co}_{0.8}\text{Fe}_{0.2}\text{O}_{3-\delta}$ (BSCF) 70 wt% + SDC 30 wt% as the cathode¹⁴. Here, an additional $\sim 10\text{ }\mu\text{m}$ of a porous layer of Ru + CeO_2 , $\sim 7\text{ wt\%}$ Ru (crystalline, $\sim 500\text{-nm}$ particles) physically mixed with CeO_2 , was coated onto the anode surface (with the exception of a small centre region for making electrical contact) to enhance catalysis of propane partial oxidation at lower temperatures. Silver net and gold wires were attached to the electrode surfaces for current collection. The cell so fabricated (Fig. 1a) was placed in a quartz tube reactor for thermal and electrochemical evaluation (Fig. 1b). We first investigated the feasibility of thermally self-sustained micro-SOFCs by measuring the thermal behaviour of anode-supported structures, both with and without the Ru + CeO_2 catalyst layer. Half cells (thin ceria electrolyte supported on porous Ni + SDC) were placed in a furnace, parallel to the gas flow direction; the cell temperature was monitored using a thermocouple protected with a thin-walled (0.4 mm) quartz tube and placed in direct contact with the centre of the anode surface. The feed gas was composed of 40 ml min^{-1} C_3H_8 , 90 ml min^{-1} O_2 and 360 ml min^{-1} He (all at standard temperature and pressure, STP, 298 K, 1 atm); this mixture has a fuel-to-oxygen ratio found elsewhere¹⁴ to be optimal for high fuel-cell power output, and an inert-to-oxygen ratio similar to air, so it is representative of fuel–air mixtures envisioned for practical applications. The furnace temperature was raised or lowered in 25°C increments, and the cell temperature recorded at each step after both the furnace and cell reached steady-state conditions.

Ignition of the partial oxidation reaction for the cell containing no additional catalyst layer occurred only when the furnace temperature was raised to 461°C ($T_{\text{cell}} = 680^\circ\text{C}$); see Fig. 2a. The reaction was sustained to somewhat lower temperatures on cooling, but was nevertheless extinguished below a furnace temperature of 361°C ($T_{\text{cell}} = 621^\circ\text{C}$). The application of a Ru + CeO_2 catalyst layer had a significant impact on the thermal behaviour of the cells; see Fig. 2a. Initiation of the partial oxidation reaction occurred at a furnace temperature of 311°C (cell temperature, 618°C), as compared to 461°C without the catalyst. Moreover, upon cooling the furnace to room temperature, partial oxidation was not extinguished even in the absence of external heating. The cell maintained a temperature of 504°C with the furnace turned off and left in the open position. When the reactor was better insulated by covering with a thermal blanket, the cell temperature increased to around 580°C . This temperature was maintained with excellent stability over the course of a 200-h test (Fig. 2b). An existing model of heat-recirculating reactors¹⁵ including heat losses to ambient and heat conduction along the tube walls was adapted to simulate the current apparatus. Model results show that the heat recirculation contributes about 280°C of temperature rise. Thus, heat recirculation, along with thermal insulation and the Ru + CeO_2 catalyst, which is highly active at low temperatures, are all important ingredients for thermal sustainability in these small fuel cells (Supplementary Fig. S1). The temperature of a two-cell stack (described further below) exposed to similar conditions was $\sim 575^\circ\text{C}$, almost identical to that of the single cell.

The difference in thermal behaviour between the untreated and Ru + CeO_2 -treated half-cells can be understood in terms of the catalytic activity of Ni + SDC versus that of Ru + CeO_2 towards propane oxidation (Fig. 3), as measured using powder samples in a standard catalytic reactor¹⁶. At temperatures below 400°C (Fig. 3a, c), Ni + SDC is essentially inert, whereas Ru + CeO_2 becomes active at temperatures just above 300°C (Fig. 3b, d). This greater degree of activity towards propane oxidation results in greater heat release at low temperatures ($300\text{--}500^\circ\text{C}$; Fig. 3e), as calculated from the propane conversion data (Fig. 3c, d) and the known enthalpies of the partial and total oxidation reactions (with total oxidation releasing more heat than partial oxidation). It is this heat release that sustains the Ru + CeO_2 -treated cells at a temperature ($500\text{--}580^\circ\text{C}$) suitable for

fuel-cell operation in the absence of external heating. Furthermore, while both catalysts produce, at temperatures below 650°C , greater quantities of the deep oxidation products (H_2O and CO_2), than the partial oxidation products (H_2 and CO), Ru + CeO_2 shows somewhat better selectivity for partial oxidation at the lower temperatures (Fig. 3f). This feature is an additional advantage of Ru + CeO_2 for fuel-cell operation.

A complete fuel-cell structure with BSCF + SDC as the cathode and a Ru + CeO_2 -treated anode was then examined for power output under thermally self-sustaining conditions. A $\text{C}_3\text{H}_8 + \text{O}_2 + \text{He}$ gas mixture with a fixed volumetric ratio of 4:9:36 and a total flow rate of $368\text{--}490\text{ ml min}^{-1}$ (at STP) was supplied to the fuel cell. Start-up was initiated by placing the cell directly into a preheated furnace (500°C). The cell temperature rose rapidly, and after a short equilibration, the fuel cell was removed from the furnace and the thermal blanket applied. In the self-sustaining mode, an open circuit voltage (OCV) of $\sim 0.7\text{ V}$ was obtained with peak power densities of 182 to 247 mW cm^{-2} and short-circuit current densities of $0.8\text{--}1.1\text{ A cm}^{-2}$ (Fig. 4). The latter two properties increased with increasing propane flow rate, although the OCV remained roughly constant (Fig. 4a). The increase in power output was accompanied by an increase of the fuel-cell temperature, from 535 to 580°C for a corresponding propane flow rate increase from 32 to 40 ml min^{-1} . The OCVs measured here are lower than can be obtained from conventional dual-chamber fuel cells, but are typical of SCFCs in which imperfect selectivity of the electrodes results in some fuel oxidation at the cathode¹⁴. Furthermore, the power densities are lower than reported previously for identical fuel cells because of the

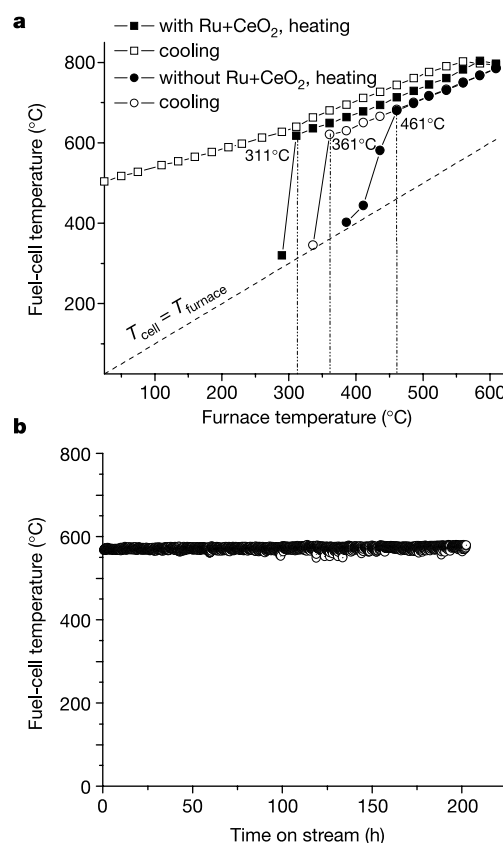


Figure 2 | The temperature recorded for the half-cells, both with and without a Ru + CeO_2 catalyst layer. The feed gas composition was 40 ml min^{-1} $\text{C}_3\text{H}_8 + 90\text{ ml min}^{-1}$ $\text{O}_2 + 360\text{ ml min}^{-1}$ He (all at STP). **a**, Steady-state fuel-cell temperature as a function of furnace temperature; **b**, temperature of the self-sustaining Ru + CeO_2 -treated cell, showing excellent long-term stability.

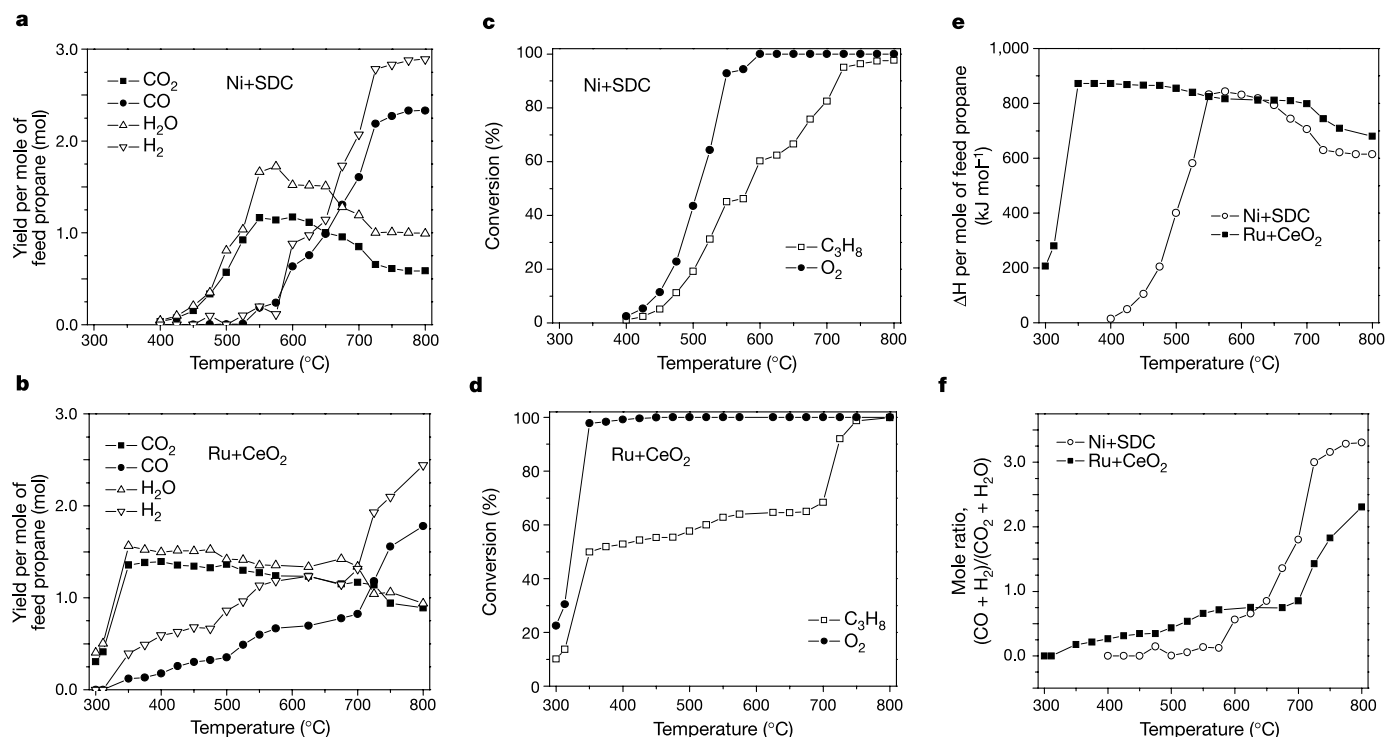


Figure 3 | Catalytic properties of Ni + SDC and Ru + CeO₂ towards propane oxidation. The feed gas composition was 10 ml min⁻¹ C₃H₈ + 22.5 ml min⁻¹ O₂ + 90 ml min⁻¹ He (all at STP). In all cases, the temperature reported is the real catalyst temperature, monitored by an in situ K-type thermocouple. **a**, Oxidation products as a function of temperature over 0.18 g of Ni + SDC powder (calcined at 1,350 °C). **b**, Oxidation products as a function of temperature over 0.18 g Ru + CeO₂ (Ru 7 wt%) catalyst powder reduced and calcined at 950 °C. **c**, **d**, Corresponding temperature dependences of propane and oxygen conversion over the

Ni + SDC and Ru + CeO₂ catalysts, respectively. **e**, Heat released as a function of temperature over the Ni + SDC and Ru + CeO₂ catalysts owing to the exothermic oxidation reactions, as calculated from the measured catalytic activity. **f**, The ratio of partial oxidation products (CO + H₂) to deep oxidation products (CO₂ + H₂O) over Ni + SDC catalyst and Ru + CeO₂ catalyst at various temperatures. Overall, Ru + CeO₂ releases more heat and is more selective for partial oxidation products at temperatures below ~600 °C than Ni + SDC.

lower fuel-cell temperature¹⁴ (which results in higher electrode and electrolyte resistances), but the values are nonetheless higher than those attained from the most advanced DMFCs, typically 100–200 mW cm⁻² (refs 17, 18). Quite significant is the fact that the electrolyte resistance, which is highly dependent on grain size and impurity content (Supplementary Fig. S2), was responsible for ~65% of the fuel-cell polarization losses (Supplementary Fig. S3). Thus, careful control of the fabrication procedures to decrease the electrolyte resistance and improved design of reactor to improve OCV can be anticipated to increase the power output by at least a factor of two. Additional design efforts are also required to improve the overall fuel utilization and thereby the efficiency, which is presently estimated at ~1%.

For practical applications, a voltage greater than 0.70 V is generally required and can be achieved by combining individual fuel cells into a stack. In conventional dual-chamber fuel cells, stacks are arranged in a cathode-facing-anode multilayer configuration, with interconnect layers separating anode and cathode chambers. Here we implement an anode-facing-anode stack configuration. Because the two anodes share the same gas, no interconnect layer is necessary, fabrication is exceedingly simple and the stack can be made very compactly. The stack produced approximately double the voltage (1.44 V) and more than double the power of a single cell (Fig. 4b), enough to operate a 1.5-V MP3 player (Supplementary Fig. S4). The 11% increase in average peak power density from 247 to 275 mW cm⁻² attained in the stacked configuration is due to the favourable gas transport pathways in the anode-facing-anode stack. Here, the partial oxidation products H₂ and CO are hindered from readily transporting to the cathode. In contrast, an anode-facing-

cathode dual-cell stack constructed without an interconnect layer resulted in near-zero OCV. In this configuration the highly reactive partial oxidation products (CO + H₂) are rapidly transported to the cathode, at which chemical oxidation is presumably catalysed. Undesirable gas-phase transport between electrodes apparently also limits the power outputs of fuel cells configured with anode and cathode strips placed on the same side of a supporting electrolyte¹⁹.

Hydrocarbon-fuelled fuel cells often suffer from carbon deposition at the anode, especially under open-circuit conditions²⁰. Because of the presence of oxygen in the fuel stream, however, SCFCs are much less susceptible to carbon coking (Supplementary Fig. S5). Our Ru + CeO₂-treated half-cell was entirely free of carbon deposits after the 200-h exposure to C₃H₈ + O₂ + He (at essentially OCV conditions). In contrast, a small amount of visible carbon was evident at the anode of our untreated half-cell after only ~24 h of exposure ($T_{\text{cell}} \approx 600$ °C). Not surprisingly, serious carbon coking, which completely destroyed the fuel cell, occurred when the Ni anode was exposed to pure propane (Supplementary Fig. S6). In the case of the Ru + CeO₂-treated cell, no carbon deposition occurred on the Ni beneath the catalyst layer (we peeled off the top layer to inspect the anode), suggesting that Ru + CeO₂ is so active that the partial oxidation was completed within this exterior layer and no propane, in fact, reached the inner Ni + SDC anode.

The ability of the fuel cell to withstand rapid thermal cycling was tested by directly inserting and removing the fuel-cell reactor from the high-temperature furnace (500 °C) multiple times. The fuel cell did not exhibit any mechanical cracks or deterioration in performance. This behaviour is achievable because the otherwise delicate seal of a conventional SOFC has been eliminated, because a thin-film

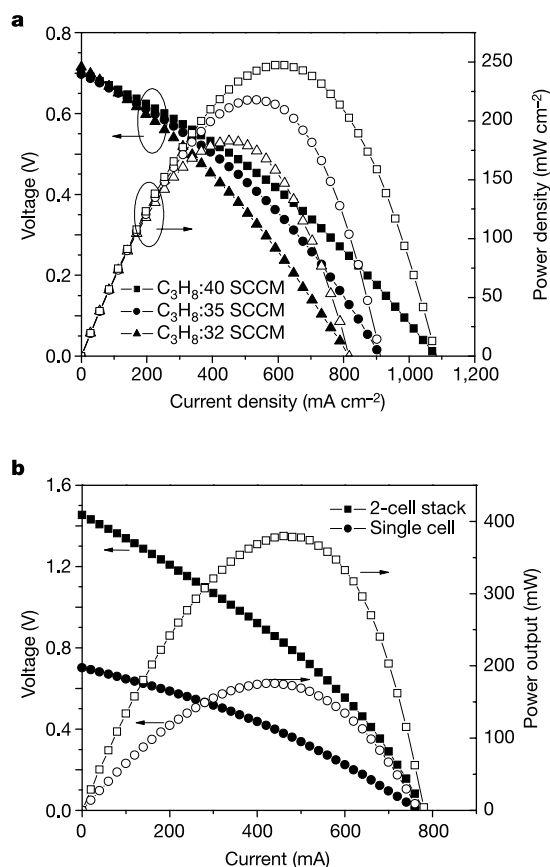


Figure 4 | Performance of the thermally self-sustained fuel cells. **a**, The cell voltage and power output versus current for a single cell at different propane feed rates, but fixed propane to oxygen to helium ratio of 4:9:36; **b**, A comparison of the performance of an anode-facing-anode, two-cell stack and a single cell at a flow composition of 40 ml min⁻¹ C₃H₈ + 90 ml min⁻¹ O₂ + 360 ml min⁻¹ He (all at STP), the cathode surface area for each cell is 0.71 cm².

electrolyte has been implemented, and because the operation temperature has been reduced to 500–600 °C. The time from cold-start to stable power output was typically less than 1 min, a competitive timescale for portable power. Although an external heat source has been employed in the present design for the start-up process, ultimately, self-initiation of the power generator may be possible using improved catalysts and a fuel that ignites on a catalyst at low temperature such as methanol, hydrogen or di-methyl ether. Several technical hurdles remain and the inherently low efficiency of the single chamber design poses additional challenges, but these results demonstrate the viability of high-energy-density hydrocarbon fuels for portable power applications.

Received 24 November 2004; accepted 23 March 2005.

- Heinzel, A., Hebling, C., Muller, M., Zedda, M. & Muller, C. Fuel cells for low power applications. *J. Power Sources* **105**, 250–255 (2002).
- Dyer, C. K. Fuel cells for portable applications. *J. Power Sources* **106**, 31–34 (2002).
- McGrath, K. M., Prakash, G. K. S. & Olah, G. A. Direct methanol fuel cells. *J. Ind. Eng. Chem.* **10**, 1063–1080 (2004).
- Perry Murray, E., Tsai, T. & Barnett, S. A. A direct-methane fuel cell with a ceria-based anode. *Nature* **400**, 649–651 (1999).
- Park, S., Vohs, J. M. & Gorte, R. J. Direct oxidation of hydrocarbons in a solid-oxide fuel cell. *Nature* **404**, 265–267 (2000).
- Tao, S. W. & Irvine, J. T. S. A redox-stable efficient anode for solid-oxide fuel cells. *Nature Mater.* **2**, 320–323 (2003).
- Dyer, C. K. A novel thin-film electrochemical device for energy conversion. *Nature* **343**, 547–548 (1990).
- Hibino, T. et al. A low-operating-temperature solid oxide fuel cell in hydrocarbon-air mixtures. *Science* **288**, 2031–2033 (2000).
- Liu, H., Ramnarayanan, R. & Logan, B. E. Production of electricity during wastewater treatment using a single chamber microbial fuel cell. *Environ. Sci. Technol.* **38**, 2281–2285 (2004).
- Stefan, I. C., Jacobson, C. P., Visco, S. J. & De Jonghe, L. C. Single chamber fuel cells: flow geometry, rate, and composition considerations. *Electrochem. Solid-State Lett.* **7**, A198–A200 (2004).
- Suzuki, T., Jasinski, P., Petrovsky, V., Anderson, H. U. & Dogan, F. Anode supported single chamber solid oxide fuel cell in CH₄-air mixture. *J. Electrochem. Soc.* **151**, A1473–A1476 (2004).
- Hibino, T. et al. A solid oxide fuel cell using an exothermic reaction as the heat source. *J. Electrochem. Soc.* **148**, A544–A549 (2001).
- Napporn, T. W., Morin, F. & Meunier, M. Evaluation of the actual working temperature of a single-chamber SOFC. *Electrochem. Solid-State Lett.* **7**, A60–A62 (2004).
- Shao, Z. P. & Haile, S. M. A High-performance cathode for the next generation solid-oxide fuel cells. *Nature* **431**, 170–173 (2004).
- Ronney, P. D. Analysis of non-adiabatic heat-recirculating combustors. *Combust. Flame* **135**, 421–439 (2003).
- Shao, Z. P., Kwak, C. & Haile, S. M. Anode-supported thin-film fuel cells operated in a single chamber configuration. *Solid State Ionics* **175**, 39–46 (2004).
- Shukla, A. K., Jackson, C. L., Scott, K. & Raman, R. K. An improved-performance liquid-feed solid-polymer-electrolyte direct methanol fuel cell operating at near-ambient conditions. *Electrochem. Acta* **47**, 3401–3407 (2002).
- Thomas, S. C., Ren, X. M., Gottesfeld, S. & Zelenay, P. Direct methanol fuel cells: progress in cell performance and cathode research. *Electrochem. Acta* **47**, 3741–3748 (2002).
- Hibino, T., Ushiki, K. & Kuwahara, Y. New concept for simplifying SOFC system. *Solid State Ionics* **91**, 69–74 (1996).
- Koh, J. H., Kang, B. S., Lim, H. C. & Yoo, Y. S. Thermodynamic analysis of carbon deposition and electrochemical oxidation of methane for SOFC anodes. *Electrochem. Solid-State Lett.* **4**, A12–A15 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Goodwin and Y. Hao of Caltech for discussions. This work is funded by the Defense Advanced Research Projects Agency, Microsystems Technology Office. Additional support has been provided by the National Science Foundation through the Caltech Center for the Science and Engineering of Materials.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.M.H. (smhaile@caltech.edu).

Lightning-induced intensification of the ionospheric sporadic E layer

C. J. Davis¹ & C. G. Johnson¹

A connection between thunderstorms and the ionosphere has been hypothesized since the mid-1920s¹. Several mechanisms have been proposed to explain this connection^{2–7}, and evidence from modelling⁸ as well as various types of measurements^{9–14} demonstrate that lightning can interact with the lower ionosphere. It has been proposed, on the basis of a few observed events¹⁵, that the ionospheric ‘sporadic E’ layer—transient, localized patches of relatively high electron density in the mid-ionosphere E layer, which significantly affect radio-wave propagation—can be modulated by thunderstorms, but a more formal statistical analysis is still needed. Here we identify a statistically significant intensification and descent in altitude of the mid-latitude sporadic E layer directly above thunderstorms. Because no ionospheric response to low-pressure systems without lightning is detected, we conclude that this localized intensification of the sporadic E layer can be attributed to lightning. We suggest that the co-location of lightning and ionospheric enhancement can be explained by either vertically propagating gravity waves that transfer energy from the site of lightning into the ionosphere, or vertical electrical discharge, or by a combination of these two mechanisms.

The Arrival Time Difference system of the UK Meteorological Office¹⁶ was used to identify lightning that occurred within the field of view of the ionospheric monitoring station at Chilton, UK, between 1993 and 2003. The system identifies broadband radio noise emitted by lightning discharges known as sferics, and is able to estimate the position of lightning to within 5 km over the UK. Hourly ionospheric measurements are made using a digital ionospheric sounder (digisonde)¹⁷ from which a series of short-wave radio pulses are transmitted (typically 1–15 MHz in 100-kHz steps). Reflection occurs when the plasma frequency of the ionospheric electrons matches the radio frequency. The plasma frequency, f_p (in MHz) is related to the number density of the electrons, N_e (in m^{-3}), by the formula $f_p = 8.98\sqrt{N_e}$.

The most common cause of sporadic E (E_s) ionization at mid-latitudes is through meteoric metal atoms that are deposited over a broad range of heights around 100 km (refs 18, 19). The small fraction of these atoms that become ionized are relatively long-lived¹⁹ (several hours), so it is possible for them to be concentrated by tidal forces and wind shears into thin, dense layers¹⁸. The localized nature of the meteor showers means that these layers are often patchy. The intensity of an E_s layer is parameterized in two ways: by the peak plasma frequency (ordinary mode), f_{oE_s} , which is a measure of the densest patches of ionization within the layer, and the blanketing frequency, f_{bE_s} , which corresponds to the lowest frequency that can penetrate the layer and is therefore a measure of the electron number density in the weakest patches. The height of an E_s layer (obtained from the flight time of the radio signal, assuming it is travelling through free-space), h'_{E_s} , is inferred from the time-of-flight of the radio pulse. This apparent or ‘virtual’ height can be artificially

increased by up to about 2 km if underlying ionization delays the propagation of the radio pulse.

For the period of the current study, 3,874 hours were identified in each of which there was at least one lightning event. The average behaviour of three ionospheric parameters was investigated for 160 hours either side of the lightning events in a superposed epoch analysis. This technique is frequently used with geophysical data to identify repeatable but weak responses to a given phenomenon^{20,21}. E_s ionization exhibits strong seasonal and diurnal trends¹⁸, which are likely to be much larger than any effect due to lightning. These trends

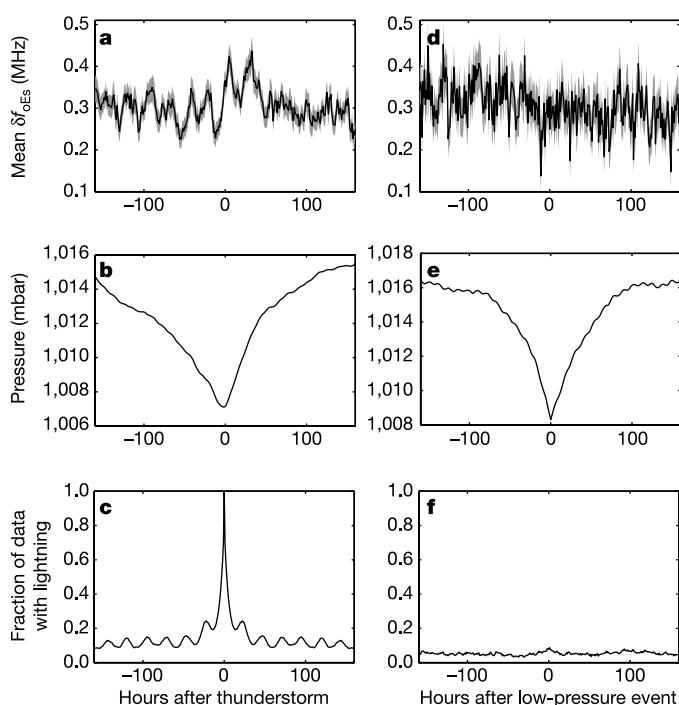


Figure 1 | Superposed epoch analyses for lightning and low-pressure events. Mean δf_{oE_s} values are shown in **a** and **d** as a black line, with the standard error in these mean values (calculated from the standard deviation of each mean divided by the square root of the number of points in that mean) represented by the width of the grey shaded area about this line. When calculating δf_{oE_s} , it is desirable to subtract the 30-day median rather than the 30-day mean, because it is less sensitive to outliers. The resulting δf_{oE_s} values are positive because the median is less than the mean for the f_{oE_s} distribution, because the distribution has large positive outliers. The average drop in pressure shown in **b** and **e** is of the order of 7 mbar (corrected to sea level). The fraction of data containing lightning in each hourly bin is shown in **c** and **f**. The number of hours containing lightning data was restricted for each pressure event, resulting in a flat distribution of lightning events (**f**).

¹Rutherford Appleton Laboratory, Chilton, Oxfordshire OX11 0QX, UK.

would dominate the study, especially as the lightning events are not randomly distributed in time (most lightning occurs on summer afternoons). To minimize the effects of these trends, therefore, the parameters δf_{oEs} , δf_{bEs} and $\delta h'_{Es}$ were defined. These are the difference between any hourly values of f_{oEs} , f_{bEs} and h'_{Es} and the median value for that hour (calculated from 30 days of data around each data point). In this way, the method becomes sensitive to deviations from the diurnal and seasonal trends rather than to the trends themselves.

Figure 1a shows the average response in δf_{oEs} to lightning. An enhancement in δf_{oEs} is seen 6 hours after lightning and again after 30 hours (the exact time of the ionospheric response may be masked by the residual diurnal variation in the data). Relative to the null hypothesis, these enhancements are significant at the 98.1% and 99.65% levels, respectively (that is, the probability that these enhancements arose by chance are 0.019 and 0.0035, respectively). The average atmospheric pressure associated with the lightning (Fig. 1b) decreases by 7 mbar. Figure 1c displays the fraction of data containing lightning in each hourly bin. Thunderstorms rarely occur in isolation, and although the data have been arranged so that there will always be a lightning strike at time zero, additional strikes may occur either side of this. A 24-hour recurrence of thunderstorms can clearly be seen in Fig. 1c.

Some previous studies have investigated the relationship between air pressure and E_s enhancements, with conflicting results^{4,5,22,23}. To test whether E_s enhancements seen over Chilton are caused by the passage of low-pressure systems, 1,331 such events were identified in atmospheric pressure readings taken at Chilton. The subsequent response of E_s was then studied in a second superposed epoch analysis. To ensure that the effects of lightning did not dominate any ionospheric response to the pressure events, the number of lightning strikes coincident with each pressure event was restricted to fewer than 29 within 64 hours of each event. No significant response in δf_{oEs} (Fig. 1d) is seen as a result of low-pressure events, despite the average pressure variation (Fig. 1e) closely matching that associated with lightning (Fig. 1c). Wilson¹ speculated that the ionospheric electric field generated by a large rain cloud may be sufficient to induce ionization at ionospheric altitudes without lightning. The fact that there is no noticeable enhancement in E_s for low-pressure events without lightning is evidence that lightning is necessary to enhance the E_s layer significantly. It is interesting to note that the residual diurnal variation in δf_{oEs} (Fig. 1d) is not as strong as for the E_s

response to lightning (Fig. 1a). Even when the pressure events in Fig. 1e were limited to times with a distribution similar to that of thunderstorm occurrence (most occurring during summer afternoons), no obvious diurnal component in δf_{oEs} emerged, despite the non-uniform diurnal distribution of pressure events. It is therefore reasonable to conclude that the diurnal variation seen in Fig. 1a is, at least in part, caused by the 24-hour recurrence of thunderstorms within the data (Fig. 1c) rather than by pressure.

Although a significant ionospheric response to lightning has been found, such an average response could be generated by a small and consistent response or by a few large and dominant responses. To distinguish between these possibilities, the distribution of δf_{oEs} values in the two days before each lightning event was compared with that seen in the two days afterwards. Any difference between these two (approximately gaussian) distributions represents a change due to lightning. This difference shows up well if the pre-lightning distribution is subtracted from the post-lightning distribution as shown in Fig. 2. Any bins with negative totals represent a decrease of such δf_{oEs} values correlated with the occurrence of lightning. Conversely, bins with positive totals represent an increase in occurrence of such δf_{oEs} values. Figure 2 therefore shows that previously small values of $|\delta f_{oEs}| < 0.5$ MHz are enhanced to values between 0 and 2 MHz. The total number of f_{oEs} values observed after lightning is 2.1% higher than the total number before, indicating that, although existing layers are enhanced, few sporadic E layers are generated by lightning. A mechanism for intensifying the ionization, such as wind shear or wave action, must already be present.

We next investigated changes to the blanketing frequency, δf_{bEs} , and to the height of the E_s layer, $\delta h'_{Es}$, in response to lightning. The average responses of these parameters are shown in Fig. 3, along with histograms equivalent to those in Fig. 2. There is only one enhancement in δf_{bEs} (significant at the 99.2% level), about 30 hours after the lightning stroke (Fig. 3a). There is a decrease in $\delta h'_{Es}$ (Fig. 3c) of around 1 km, significant at the 99.5% level. This may represent an underestimate of the real drop in height, as any lightning-induced ionization below the E_s layer would retard the radio pulse used to estimate the layer height.

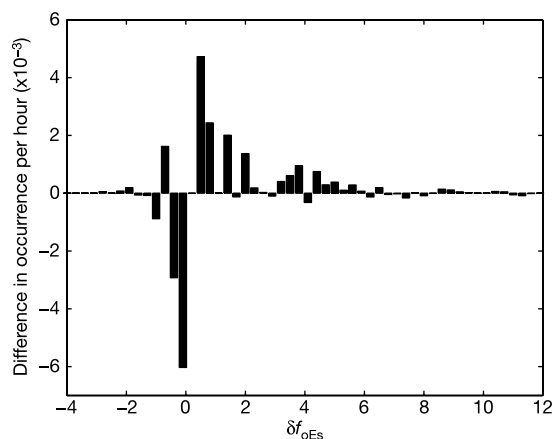


Figure 2 | Redistribution of δf_{oEs} values as a result of lightning. The distribution of δf_{oEs} values for the 48 hours before lightning events was subtracted from the distribution of δf_{oEs} values measured up to 48 hours after lightning events. The distributions were divided by the total number of hours examined (48 per thunderstorm for each distribution) to give occurrence per hour. A 2.1% rise in the occurrence of E_s after lightning results in the sum of positive bins being slightly larger than the sum of negative bins.

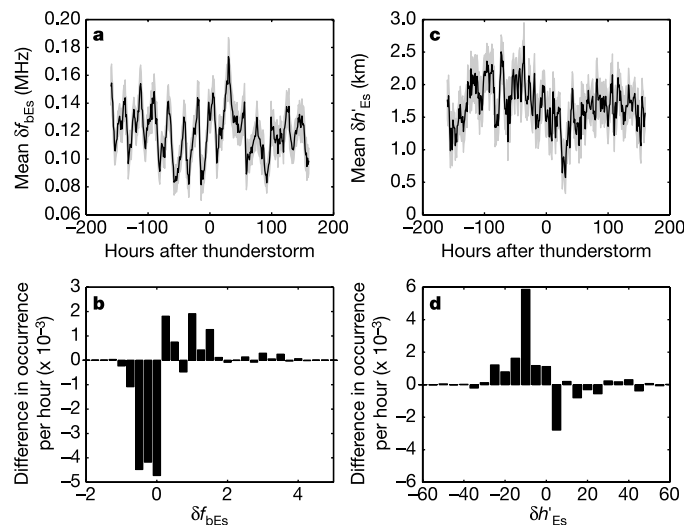


Figure 3 | Change in blanketing frequency and E_s layer height in response to lightning. The response to the same 3,874 thunderstorm events of the sporadic E blanketing frequency, δf_{bEs} (a, b), and the height of the E_s layer, $\delta h'_{Es}$ (c, d). The lines and shaded areas for a and c are the same as Fig. 1a, d but for values of δf_{bEs} and $\delta h'_{Es}$, respectively. The average δf_{bEs} value is enhanced 30 hours after lightning. b, d. Same as Fig. 2 but for changes in the distribution of δf_{bEs} and $\delta h'_{Es}$, respectively. Again, it can be seen that the average response to lightning is small but consistent, with no one event dominating either distribution.

Although gross pressure changes have been ruled out as the cause of the ionospheric enhancements seen in this study, lightning-associated wave activity could be responsible. A propagation time of six hours from cloud top to ionosphere represents a wave propagating at a velocity of around 5 m s^{-1} , which is typical for the vertical component of such a gravity wave generated in the troposphere²⁴. Mesospheric wind measurements are known to exceed modelled predictions²⁵, and this is thought to be due to gravity waves transporting energy from the troposphere.

A method for producing localized intensifications in E_s through wave action has already been proposed²⁶. If such a mechanism were solely responsible for localized enhancements in E_s through horizontal redistribution of existing ionization, no additional ionization would be created, and while δf_{oEs} would increase, δf_{bEs} may be expected to decrease as ionization from weaker patches within the layer was gathered up into the enhanced patches. No such decrease in δf_{bEs} was detected in the present study.

Because the horizontal wavelength is greater than the vertical wavelength in the mechanism²⁶ proposed above, waves modulating the ionosphere above Chilton would need to be generated in the troposphere beyond a horizontal distance of 100 km in order that the vertical component might reach ionospheric altitudes by the time the wave had reached Chilton. To find the optimum location for lightning to influence E_s above Chilton, δf_{oEs} data were binned according to the bearing and distance of the lightning. No significant direction dependence was found, and the ionospheric response was insensitive to distance for lightning within 100 km, decreasing with distance beyond 100 km. This indicates that this particular mechanism is unlikely to be responsible for the observed enhancement, although it does not rule out wave action completely. Convective instability above storm clouds can generate gravity waves that propagate vertically, and these have been observed in the mesosphere²⁷. Instabilities have also been shown to be generated in the ionosphere by lightning modulating the ambient electric field there²⁸, and there is evidence that very-low-frequency waves launched by lightning can induce further lightning²⁹.

Optical emissions known as blue jets, associated with sferics⁹, have been seen to propagate from cloud tops to an altitude of 70 km, and gravity waves associated with other upward-propagating optical features, called sprites, have been seen in the mesosphere¹⁰. Gravity waves launched from this altitude would take considerably less time to propagate 30 km vertically to the E_s layer. Measurements suggest¹⁰ that such waves propagate upwards at an angle of around 27° and would therefore reach the E_s layer within a radius of 15 km from the thunderstorm, consistent with our observations.

Measurements of ionospheric electric fields above storm clouds (typically 20 mV m^{-1})^{13,14} have been significantly less than those estimated to be required for electrical discharge in the ionosphere¹ (in excess of 48 mV m^{-1}). This calculation¹ did not consider the presence of metallic atoms, which would undoubtedly lower the electric field necessary to cause electrical discharge, because their ionization potentials are significantly below those of gas species³⁰ (for example, 5.139 eV for Na compared with 12.06 eV for O_2).

The results presented here show a statistically significant enhancement of the ionospheric E_s layer due to lightning. The localized nature of this enhancement is evidence that the mechanism involved acts vertically, either through electrical discharge or wave action, or a combination of both. The fact that few E_s layers are created as a result of lightning is evidence that a wind shear must also be present. Indeed, gravity waves launched by lightning may even act to enhance this wind shear^{29,30}.

Received 29 September 2004; accepted 6 April 2005.

1. Wilson, C. T. R. The electric field of a thundercloud and some of its effects. *Proc. Phys. Soc. Lond.* **37**, 32–37 (1925).
2. Bauer, S. J. A possible troposphere-ionosphere relationship. *J. Geophys. Res.* **62**, 425–430 (1957).

3. Bhar, J. N. & Syam, P. Effect of thunderstorms and magnetic storms on the ionisation of the Kennelly-Heaviside Layer. *Phil. Mag. J. Sci.* **23**, 513–528 (1937).
4. Mitra, S. K. & Kundu, M. R. Thunderstorms and sporadic E ionisation of the ionosphere. *Nature* **174**, 798–799 (1954).
5. Rastogi, R. G. Thunderstorms and sporadic E layer ionisation. *Indian J. Meteorol. Geophys.* **1**, 43–54 (1957).
6. Ross, W. B. & Henderson, J. T. Radio studies of the ionosphere. *Nature* **133**, 523–524 (1933).
7. Velinov, P. I., Spassov, Chr. W. & Kolev, S. I. Ionospheric effects of lightning during the increasing part of solar cycle 22. *J. Atmos. Terr. Phys.* **54**(10), 1347–1353 (1992).
8. Taranenko, Y. N., Inan, U. S. & Bell, T. F. Interaction with the lower ionosphere of electromagnetic pulses from lightning: heating, attachment and ionisation. *Geophys. Res. Lett.* **20**, 1539–1542 (1993).
9. Pasko, V. P., Stanley, M. A., Mathews, J. D., Inan, U. S. & Wood, T. G. Electrical discharge from a thundercloud top to the lower ionosphere. *Nature* **416**, 152–154 (2002).
10. Sentman, D. D. *et al.* Simultaneous observations of mesospheric gravity waves and sprites generated by a midwestern thunderstorm. *J. Atmos. Terr. Phys.* **65**, 537–550 (2003).
11. Inan, U. S., Rodriguez, J. V. & Idone, V. P. VLF signatures of lightning-induced heating and ionisation of the night-time D-region. *Geophys. Res. Lett.* **20**, 2355–2358 (1993).
12. Inan, U. S., Slingeland, A. & Pasko, V. P. VLF and LF signatures of mesospheric/lower ionospheric response to lightning discharges. *J. Geophys. Res.* **101**, 5219–5238 (1996).
13. Kelley, M. C. *et al.* Electrical measurements in the atmosphere and the ionosphere over an active thunderstorm. 1. Campaign overview and initial ionospheric results. *J. Geophys. Res.* **90**, 9815–9823 (1985).
14. Kelley, M. C., Ding, J. G. & Holzworth, R. H. Intense ionospheric electric and magnetic field pulses generated by lightning. *Geophys. Res. Lett.* **17**, 2221–2224 (1990).
15. Watson-Watt, R. A. Discussion on the ionosphere. *Proc. R. Soc.* **141**, 715–718 (1933).
16. Lee, A. C. L. Ground truth confirmation and theoretical limits of an experimental VLF arrival time difference lightning flash location system. *Q. J. R. Meteorol. Soc.* **115**, 1147–1166 (1989).
17. Bibl, K. & Reinisch, B. W. The universal digital ionosonde. *Radio Sci.* **13**, 519–530 (1978).
18. Whitehead, J. D. Recent work on mid-latitude and equatorial sporadic-E. *J. Atmos. Terr. Phys.* **51**, 401–424 (1989).
19. Plane, J. M. C., Self, D. E., Vondrak, T. & Woodcock, K. R. I. Laboratory studies and modelling of mesospheric iron chemistry. *Adv. Space Res.* **32**, 699–708 (2003).
20. Davis, C. J., Wild, M. N., Lockwood, M. & Tulunay, Y. K. Ionospheric and geomagnetic responses to changes in IMF Bz: a superposed epoch study. *Ann. Geophys.* **15**, 217–230 (1997).
21. Samson, J. C. & Yeung, K. L. Some generalizations on the method of superposed epoch analysis. *Planet. Space Sci.* **34**, 1133–1142 (1986).
22. Shatkhin, Kh. J. Ionospheric-tropospheric relationships. *Geomag. Aeronomy* **4**, 623–625 (1964).
23. Shrestha, K. L. Sporadic-E and atmospheric pressure waves. *J. Atmos. Terr. Phys.* **33**, 205–211 (1971).
24. Liu, J. Y. *et al.* Vertical phase and group velocities of internal gravity waves derived from ionograms during the solar eclipse of 24 October 1995. *J. Atmos. Solar-Terr. Phys.* **60**, 1679–1686 (1998).
25. Larsen, M. F. Wind shears in the mesosphere and lower thermosphere: Results from four decades of chemical release wind measurements. *J. Geophys. Res.* **107**(A8), 1–14 (2002).
26. Chimonas, G. Enhancement of sporadic E by horizontal transport within the layer. *J. Geophys. Res.* **76**, 4578–4586 (1971).
27. Sato, K. Vertical wind disturbances in the afternoon of mid-summer revealed by the MU-radar. *Geophys. Res. Lett.* **19**, 1943–1946 (1992).
28. Woodman, R. F. & Kudeki, E. A causal relationship between lightning and explosive spread F. *Geophys. Res. Lett.* **11**, 1165–1167 (1984).
29. Armstrong, W. C. Lightning triggered from the Earth's magnetosphere as the source of synchronised whistlers. *Nature* **327**, 405–408 (1987).
30. Schunk, R. W. & Nagy, A. F. *Ionospheres: Physics, Plasma Physics, and Chemistry* 243 (Cambridge Univ. Press, Cambridge, 2000).

Acknowledgements The authors wish to thank S. Hockenhull for prompting this investigation, J. Smith for maintaining the Chilton ionosonde, P. Odams of the UK Meteorological Office for supplying the Arrival Time Difference data, A. Ruddell for supplying the pressure data and E. Williams for his support and encouragement. The ionosonde at Chilton is funded by the UK Particle Physics and Astronomy Research Council, as is the UK Solar System Data Centre at the Rutherford Appleton Laboratory, where ionospheric data from Chilton and other stations are made available.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.J.D. (C.J.Davis@rl.ac.uk).

LETTERS

Geomagnetic dipole strength and reversal rate over the past two million years

Jean-Pierre Valet¹, Laure Meynadier² & Yohan Guyodo³

Independent records of relative magnetic palaeointensity from sediment cores in different areas of the world can be stacked together to extract the evolution of the geomagnetic dipole moment^{1,2} and thus provide information regarding the processes governing the geodynamo. So far, this procedure has been limited to the past 800,000 years (800 kyr; ref. 3), which does not include any geomagnetic reversals. Here we present a composite curve that shows the evolution of the dipole moment during the past two million years. This reconstruction is in good agreement with the absolute dipole moments derived from volcanic lavas, which were used for calibration. We show that, at least during this period, the time-averaged field was higher during periods without reversals but the amplitude of the short-term oscillations remained the same. As a consequence, few intervals of very low intensity, and thus fewer instabilities, are expected during periods with a strong average dipole moment, whereas more excursions and reversals are expected during periods of weak field intensity. We also observe that the axial dipole begins to decay 60–80 kyr before reversals, but rebuilds itself in the opposite direction in only a few thousand years.

Stacks of globally distributed relative palaeointensity records provide the best source of information regarding the evolution of the relative strength of the dipole field in the past. The remarkable consistency of the first stacks (Sint-200 and Sint-800) with alternative data sets derived from near bottom sea-floor magnetic anomalies^{4,5} and the good agreement with cosmogenic isotope records^{6,7} validated their geomagnetic origin. With the accumulation of recent studies from different oceanic basins, it has become possible to perform a first global stack for the past 2 Myr in order to document the evolution of the geomagnetic field during periods with field reversals. The most recent database^{8–21} of relative palaeointensity incorporates 15 records between 0.6 and 1.5 Myr ago. All display the same main features, but some discrepancies are caused by subtle changes in the response function of the sediment²² and by inaccuracies in the depth-versus-time correspondence assigned to each sediment core. In fact, it has been shown^{23–25} that very few variations shorter than 20 kyr can be extracted with good confidence from composite curves. Guyodo and Valet²⁵ recently correlated the intensity minima of 15 records spanning the time interval 0.75–1.25 Myr ago using the ODP site 983 (ref. 11) record as a target for the common timescale. The stack derived from this data set was calculated after normalizing the signals of all 15 records to both equal mean and variance over the interval common to all records. However, the variability of the signal decreases with temporal resolution. For this reason, it is difficult to combine different stacks obtained from different cores over different periods. In order to produce a new global stack for the time interval 0.6–2 Myr ago and then to be able to combine it with Sint-800, we thus restricted the present study to records with similar temporal resolution.

We selected 10 out of 15 records that were obtained with similar resolution and displayed coherent features. Records that were either poorly constrained in time or probably contaminated by lithological variations were not considered. Except for some intervals, core KS752 (ref. 8) is of rather poor quality and was not retained. Cores LC07 (ref. 9) and KK78 (ref. 10) show a very large increase of palaeointensity between 1.2 and 1.8 kyr ago, which is not present in the other records. Last, the records from ODP sites 983 (ref. 11) and 1101 (ref. 12) do not go beyond 1.1 Myr ago. In addition, they were obtained with much higher resolution than the rest of the database, and may have climatically contaminated intervals²⁶.

Before any calculation, the records were resampled at 1 kyr intervals. The record from ODP site 851, which covers the entire 2-Myr-long interval with relatively constant deposition rate, was selected as a target for establishing the common timescale. The influence of the

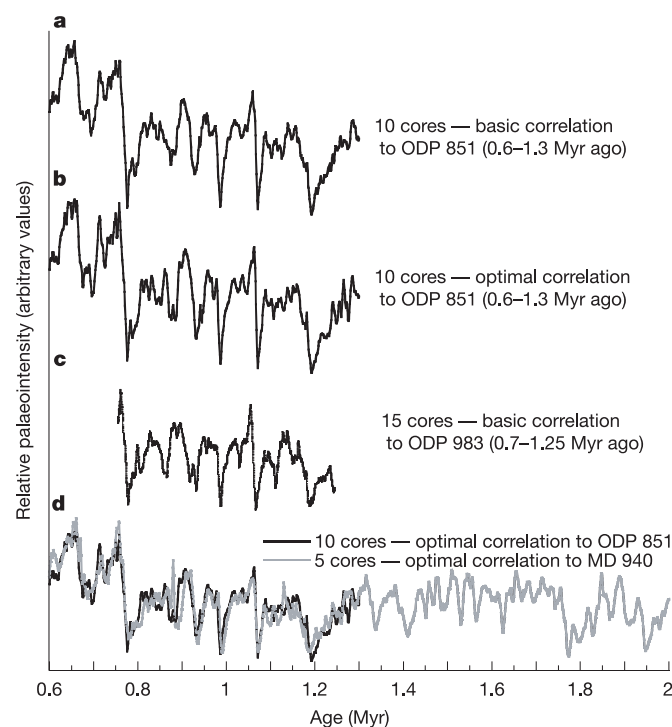


Figure 1 | Various stacks obtained for the period 0.6–2 Myr ago.

a, b, Obtained by matching a subset of 10 cores tied to the reversals (**a**) or tied to reversals and intensity minima (**b**). **c**, Obtained using all 15 records. **d**, Obtained with 10 cores between 0.6 and 1.3 Myr ago (black curve) and 5 cores between 0.6 and 2 Myr ago (grey curve).

¹Géomagnétisme et Paléomagnétisme (UMR CNRS 7577), ²Géochimie et Cosmochimie (UMR CNRS 7579), Institut de Physique du Globe de Paris, 4 Place Jussieu, 75252 Paris Cedex 05, France. ³Laboratoire des Sciences du Climat et de l'Environnement, Avenue de la Terrasse, 91190 Gif-sur-Yvette, France.

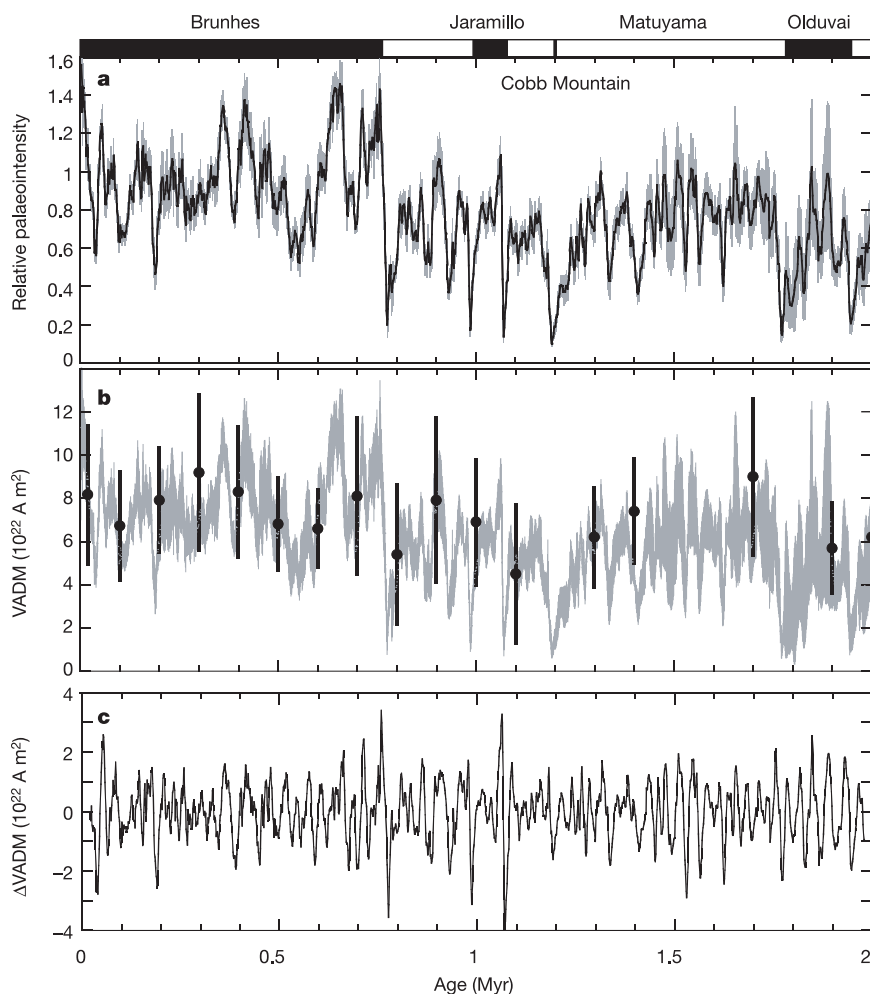


Figure 2 | Composite Sint-2000 record depicting the evolution of the field intensity during the past 2 Myr. **a**, Curve obtained after stacking all individual records, shown with the 95% confidence interval in grey. The succession of the polarity intervals is depicted by the black (white) bars for normal (reverse) polarity at the top of the panel. **b**, Calibration of Sint-2000

using the absolute palaeointensity from volcanic records converted into virtual axial dipole moments (VADMs). The VADMs were averaged over successive time intervals (at least 0.1 Myr long). Error bars indicate the dispersion of the VADMs. **c**, Fluctuations of the dipole moment around its

correlations between cores was tested for the period 0.6–1.3 Myr ago, which incorporates the largest number of records. The stack (Fig. 1a) calculated after correlating the intensity minima corresponding to the reversals (Brunhes–Matuyama, onset and termination of the Jaramillo subchron, Cobb Mountain and Upper Olduvai) was actually very similar to the ‘optimal’ stack (Fig. 1b) obtained after improving the correlations by matching other minima of intensity without violating the original depth–time correspondence. We noticed also that there is no major difference between the curves tuned to different targets (Fig. 1d), and that the stacks involving all 15 records²⁵ (Fig. 1c) or 10 cores only (Fig. 1a, b) do not significantly differ from each other. The oldest time period (2–1.5 Myr ago) is documented by 5 records only, but with relatively good geographical coverage. We verified that the stack calculated using this subset of 5 cores is identical to the one derived from all 10 cores over their common time interval, 0.6–1.3 Myr ago (Fig. 1d). We thus infer that it is representative of the field, but we are aware that additional records could reduce uncertainties for this period. In order to scale the 0.6–2-Myr-old new stack with Sint-800, we assigned the same mean value to their common time interval, 0.6–0.8 Myr ago.

The curve of the field intensity for the past 2 Myr shown in Fig. 2a is referred to as Sint-2000. As expected, larger dispersion is observed for the time interval 1.5–2 Myr ago because there are fewer records. Records of relative palaeointensity from sediments can be

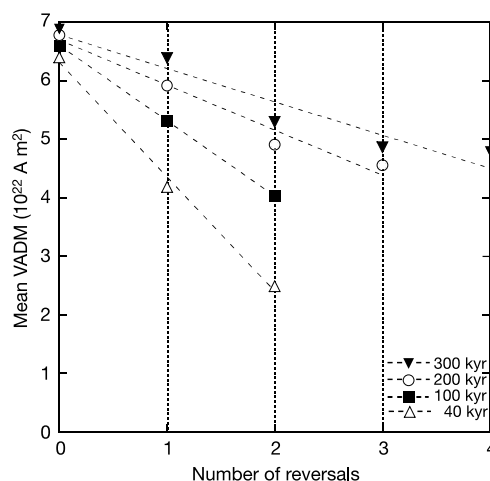


Figure 3 | Time-averaged dipole moment calculated within sliding windows of different sizes, as a function of the number of reversals contained within each window. Window sizes are 40, 100, 200 and 300 kyr. The intervals immediately surrounding the reversals have been removed from the calculation. The mean dipole moment decreases with the number of reversals. Because large windows can incorporate periods of low and high reversal frequency, the slope decreases with the size of the windows.

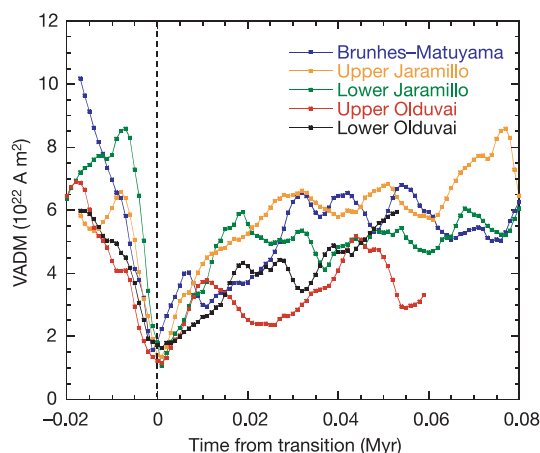


Figure 4 | Field intensity variations across the five reversals occurring during the past 2 Myr. In this figure we superimposed the changes in dipole moment during the 80 kyr and 20 kyr time intervals respectively preceding and following each reversal. Note the 60–80-kyr-long decrease preceding the reversals, and the rapid recovery following the transitions.

calibrated to absolute determinations from volcanic records. However, this procedure is delicate, because volcanic records represent instantaneous measurements of the total field and their ages are associated with relatively large uncertainties. In order to remove the contribution of non dipole components and to reduce age uncertainties, we averaged the virtual axial dipole moments (VADMs) of the 2004 updated volcanic database²⁷ over 0.1-Myr-long intervals. We then calculated the time-averaged VADM ($(7.46 \pm 1.16) \times 10^{22} \text{ A m}^2$) for the past 0.8 Myr and used this value for calibration. This interval was selected because all successive 0.1-Myr-long segments incorporate enough data points to calculate a reasonable estimate of the mean dipole moment. There is an overall satisfactory agreement between each successive time-averaged volcanic value and the calibrated record shown in Fig. 2b.

A striking characteristic of Sint-2000 is the succession of periods with different mean values. During the Brunhes chron, the dipole oscillated around a value of $(7.5 \pm 1.7) \times 10^{22} \text{ A m}^2$, which is significantly larger than during the previous 400 kyr, $(5.3 \pm 1.5) \times 10^{22} \text{ A m}^2$. It is tempting to link the existence of a low (strong) dipole with the presence (absence) of reversals like during the Matuyama (Brunhes) or during the 0.5-Myr-long interval without reversals between the Gauss–Matuyama and the Cobb Mountain reversals. To provide a more quantitative picture of field strength as a function of reversal frequency, we calculated successive running averages of the field intensity using different rectangular windows (40, 100, 200, 300 kyr). The mean VADMs within all intervals were classified with respect to the number of reversals (N) that they contain (a 100-kyr-long interval embraces a maximum of two reversals). The mean of each category gives an estimate of the field strength for the intervals containing between 0 and 4 reversals. For each class of N reversals, the average field value and its standard deviation are reported in Fig. 3 as a function of N . The results show that the averaged field intensity linearly decreases with respect to the number of reversals. As expected, the slope decreases for increasing sizes of the windows as they involve longer time-averaged variations. The Cobb event was considered as a full polarity interval with two successive reversals separated by a 10-kyr-long interval of low intensity. If we were to see it as a single transition, the slope of the plot would be larger. We also removed the data contained within the 20-kyr-long intervals surrounding the transitions, and verified that they did not affect the calculations. From these results, we conclude that at least during this period the dipole field strength appears to be a dominant factor controlling the frequency of reversals. A similar but weaker dependence of average field intensity and length of polarity

intervals was previously reported²⁸ in the study of an 11-Myr-long record of the Oligocene.

We wondered whether the variability of the dipole (ΔVADM) is sensitive to its mean value. This parameter was approximated after smoothing Sint-2000 using a running average with a 40 kyr window size and subtracting each successive value from the corresponding initial data point. The plot in Fig. 2c shows that ΔVADM remains more or less constant. Thus minima of intensity are more easily reached when the averaged moment is weak than when it is strong. A direct consequence is that the ratio of the axial dipole field intensity to the non-axial dipole part is lower, which should generate more geomagnetic instabilities during weak field periods. One may then wonder whether this is consistent with the apparently frequent occurrence of excursions (roughly one every 100 kyr) that have been reported for the present polarity interval of rather strong dipole. In fact, a reduction by only half the present-day field²⁹ is enough to generate anomalous (seen as excursions) directions covering one-fifth of the Earth's surface.

Because reversal positions were used as stratigraphic markers to correlate the records, timescale uncertainties are much smaller during these time intervals, which allows us to compare the pattern of the field variations immediately preceding and following each transition. In Fig. 4, we superimposed the field variations during these two phases for each polarity transition. The dominant feature is that each reversal was followed by a sudden and large field recovery. All curves show also that the dipole decayed before the transitions over a time interval of about 60–80 kyr. Because these features are present in sediments of various lithologies and physical properties, it is difficult to envisage that they could have been caused by post-depositional reorientation of magnetic grains^{30,31}. Note that such processes could not account for the variations observed across the Cobb mountain event (1.19 Myr ago). Indeed in this case the restoration of the initial polarity after a rapid succession of two reversals is incompatible with the coexistence of directions with opposite polarities at the same stratigraphic levels. It is also significant that similar asymmetry between the pre- and post-reversal stages has been reported in the most detailed volcanic records of reversals²⁵. If confirmed, the respective durations of these two stages suggest that diffusive processes could dominate the pre-reversal episode while induction could drive the dipole recovery. Altogether, these results bring new constraints to models for the geodynamo, and emphasize the importance of palaeointensity studies for our understanding of the geomagnetic field.

Received 23 July 2004; accepted 13 April 2005.

- Guyodo, Y. & Valet, J. P. Relative variations in geomagnetic intensity from sedimentary records: the past 200 thousand years. *Earth Planet. Sci. Lett.* **143**, 23–26 (1996).
- Laj, C., Kissel, C., Mazaud, M., Channell, J. E. T. & Beer, J. North Atlantic paleointensity stack since 75 ka (NAPIS 75) and the duration of the Laschamp event. *Phil. Trans. R. Soc. Lond. A* **358**, 1009–1025 (2000).
- Guyodo, Y. & Valet, J. P. Global changes in intensity of the Earth's magnetic field during the past 800 kyr. *Nature* **399**, 249–252 (1999).
- Gee, J., Cande, S. C., Hildebrand, J. A., Donnelly, J. A. & Parker, R. L. Geomagnetic intensity variations over the past 780 kyr obtained from near-seafloor magnetic anomalies. *Nature* **408**, 827–832 (2000).
- Pouliquen, G., Gallet, Y., Dymont, J., Patriat, P. & Tamura, C. A geomagnetic record over the last 4 million years from deep-tow magnetic profiles across the Central Indian Ridge. *J. Geophys. Res.* **106**, 10941–10960 (2001).
- Frank, M. A 200 kyr record of cosmogenic radionuclide production rate and geomagnetic field intensity from ^{10}Be in globally stacked deep-sea sediments. *Phil. Trans. R. Soc. Lond. A* **358**, 1089–1107 (2000).
- Baumgartner, S. et al. Geomagnetic modulation of the ^{36}Cl flux in the Grip ice core, Greenland. *Science* **279**, 1330–1332 (1998).
- Valet, J.-P., Meynadier, L., Bassinot, F. C. & Garnier, F. Relative paleointensity across the last geomagnetic reversal from sediments of the Atlantic, Indian and Pacific oceans. *Geophys. Res. Lett.* **21**, 485–488 (1994).
- Dinares-Turell, J., Sagnotti, L. & Roberts, A. P. Relative geomagnetic paleointensity from the Jaramillo subchron to the Matuyama/Brunhes boundary as recorded in a Mediterranean piston core. *Earth Planet. Sci. Lett.* **194**, 327–341 (2002).

10. Laj, C. & Kissel, C. Geomagnetic field intensity and reversals for the last 1.8 Ma from a central equatorial Pacific core. *Geophys. Res. Lett.* **23**, 3393–3396 (1996).
11. Channell, J. E. T. & Kleiven, H. F. Geomagnetic paleointensity and astrochronological ages for the Matuyama-Brunhes boundary and the boundaries of the Jaramillo subchron: paleomagnetic and oxygen isotope records from ODP site 983. *Phil. Trans. R. Soc. Lond. A* **358**, 1027–1047 (2000).
12. Guyodo, Y., Richter, C. & Valet, J. P. Paleointensity record from Pleistocene sediments off the California Margin. *J. Geophys. Res.* **104**, 22953–22964 (1999).
13. Guyodo, Y., Acton, G. D., Brachfeld, S. & Channell, J. E. T. A sedimentary paleomagnetic record of the Matuyama chron from the western Antarctic margin. *Earth Planet. Sci. Lett.* **191**, 61–74 (2001).
14. Hayashida, A., Verosub, K. L., Heider, F. & Leonhardt, R. Magnetostratigraphy and relative paleointensity of late Neogene sediments at ODP leg 167 site 1010 off Baja California. *Geophys. J. Int.* **139**, 829–840 (1999).
15. Horng, C. S., Roberts, A. P. & Liang, W.-T. Astronomically tuned record of relative geomagnetic paleointensity from the western Philippine sea. *J. Geophys. Res.* **108**, 2059, doi:10.1029/2001JB001698 (2003).
16. Kent, D. V. & Opdyke, N. D. Paleomagnetic field intensity recorded in a Brunhes epoch deep-sea sediment core. *Nature* **266**, 156–159 (1977).
17. Kok, Y. S. & Tauxe, L. A relative geomagnetic paleointensity stack from Ontong-Java plateau sediments for the Matuyama. *J. Geophys. Res.* **104**, 25401–25413 (1999).
18. Meynadier, L., Valet, J.-P., Bassinot, F. C., Shackleton, N. & Guyodo, Y. Asymmetrical saw-tooth pattern of the geomagnetic field intensity from equatorial sediments in the Pacific and the Indian oceans. *Earth Planet. Sci. Lett.* **126**, 109–127 (1994).
19. Sato, T., Kikuchi, H., Nakashizuka, M. & Okada, M. Quaternary geomagnetic field intensity: constant periodicity or variable period? *Geophys. Res. Lett.* **25**, 2221–2224 (1998).
20. Schneider, D. A., Kent, D. V. & Mello, G. A. A high-resolution marine sedimentary record of geomagnetic intensity during the Brunhes chron. *Earth Planet. Sci. Lett.* **111**, 395–405 (1992).
21. Valet, J.-P. & Meynadier, L. Geomagnetic field intensity and reversals during the past four million years. *Nature* **366**, 234–238 (1993).
22. Valet, J.-P. Time variations in geomagnetic intensity. *Rev. Geophys.* **41**(1), 1004 (2003).
23. Guyodo, Y. & Channell, J. E. T. Effects of variable sedimentation rates and age errors on the resolution of sedimentary paleointensity records. *Geochim. Geophys. Geosyst.* **10**, doi:10.29/2001GC000211 (2002).
24. McMillan, D. G., Constable, C. G. & Parker, R. L. Assessing the dipolar signal in stacked paleointensity records using a statistical error model and geodynamo simulations. *Physics Earth Planet. Inter.* **145**, 37–54 (2004).
25. Guyodo, Y. & Valet, J. P. A comparison of relative paleointensity records of the Matuyama chron for the period 0.75–1.25 Ma. *Phys. Earth. Planet. Inter.* (in the press).
26. Guyodo, Y., Gaillet, P. & Channell, J. E. T. Wavelet analysis of relative geomagnetic paleointensity at ODP Site 983. *Earth Planet. Sci. Lett.* **184**, 109–123 (2000).
27. Perrin, M. & Schnepf, E. IAGA paleointensity database: distribution and quality of the dataset. *Phys. Earth Planet. Inter.* **147**(2–3), 255–267 (2004).
28. Constable, C. G., Tauxe, L. & Parker, R. L. Analysis of 11 Myr of geomagnetic intensity variation. *J. Geophys. Res.* **103**, 17735–17748 (1998).
29. Quidelleur, X., Gillot, P.-Y., Carlut, J. & Courtillot, V. The age and duration of the Matuyama–Brunhes transition from new K–Ar data from La Palma (Canary Islands) and revisited $^{40}\text{Ar}/^{39}\text{Ar}$ ages. *Earth Planet. Sci. Lett.* **168**, 233–242 (1999).
30. Meynadier, L. & Valet, J.-P. Post-depositional realignment of magnetic grains and asymmetrical saw-toothed pattern of magnetization intensity. *Earth Planet. Sci. Lett.* **140**, 123–132 (1996).
31. Mazaud, A. Saw-tooth variation in magnetic intensity profiles and delayed acquisition of magnetization in deep sea cores. *Earth Planet. Sci. Lett.* **139**, 379–386 (1996).

Acknowledgements This work was supported by the INSU-CNRS programme 'Dynamique et Evolution de la Terre Interne'.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.-P.V. (valet@ipgp.jussieu.fr).

LETTERS

Effects of species and functional group loss on island ecosystem properties

David A. Wardle^{1,2} & Olle Zackrisson¹

Considerable recent attention has focused on predicting how the losses of species and functional groups influence ecosystem properties^{1,2}, but the extent to which these effects vary among ecosystems remains poorly understood^{3,4}. Island systems have considerable scope for studying how biotic and abiotic factors influence processes in different ecosystems, because they enable the simultaneous study of large numbers of independent replicate systems at ecologically meaningful spatial scales^{5–7}. We studied a group of 30 islands in northern Sweden, for which island size determined disturbance history, and therefore vegetation successional stage and biotic and abiotic ecosystem properties. On each island we conducted a seven-year study that involved experimental removals of combinations of both plant functional groups and plant species. We show that although losses of functional groups and species often impaired key ecosystem processes, these effects were highly context-dependent and strongly influenced by island size. Our study provides evidence that the consequences of biotic loss for ecosystem functioning vary greatly among ecosystems and depend on the specific abiotic and biotic attributes of the system.

The recent interest in understanding how losses of biota from real ecosystems influence ecosystem processes has resulted in a significant and growing number of studies, some of which have suggested that biodiversity is a major driver of ecosystem properties^{4,8,9} and others of which have not^{10–12}. The interpretations of many of these studies have been hotly debated^{1,2}. However, it seems likely that discrepancies between the results of different studies are attributable in large part to

the differences in context between studies. For example, whether or not ecosystem properties and functions are responsive to diversity can depend on soil nutrient availability³, temporal factors¹³, the type of ecosystem functions considered², trophic interactions¹⁴ and spatial scale¹. However, definite empirical evidence for such context dependence in real ecosystems is scarce.

In this study, we used a group of 30 forested islands in two adjacent lakes in northern Sweden, Lake Uddjaure and Lake Hornavan (65° 55' N to 66° 09' N and 17° 43' E to 17° 55' E), to compare how contrasting ecosystems respond to losses of plant functional groups and species. For these islands, the primary driver of ecosystem properties is disturbance regime: large islands burn more often than smaller ones because they have a larger area that can be intercepted by lightning strikes¹¹. This creates a successional gradient across islands; large islands that are early-successional support plant species that grow faster, support more biomass and produce litter of higher quality than do smaller islands^{11,15}. This in turn affects ecosystem properties, with smaller islands supporting less productivity, lower rates of nutrient cycling, and lower supply rates of available soil nutrients compared with larger islands^{11,15}. This system offers potential for studying how interactions between resident biota and the current state of physical and chemical properties can affect ecosystem properties. We studied 30 islands, which include ten in each of three size classes known to differ markedly in successional age and therefore in biotic and abiotic factors; these classes are 'large' (>1.0 ha), 'medium' (0.1–1.0 ha) and 'small' (<0.1 ha)¹⁵.

Table 1 | Effects of plant functional group removal, island size and two-way interactions on above- and belowground properties

	Response variable*									
	R	S	M	A	R × S	R × M	S × M	R × A	S × A	M × A
Total shrub cover	6.5 (0.011)	—	0.6 (0.446)	4.3 (0.015)	—	0.2 (0.701)	—	1.7 (0.185)	—	0.0 (0.975)
Moss biomass	0.3 (0.867)	0.1 (0.723)	—	3.7 (0.028)	0.3 (0.563)	—	—	0.0 (0.970)	0.2 (0.814)	—
BR	27.5 (<0.001)	15.6 (<0.001)	0.9 (0.347)	10.0 (<0.001)	3.2 (0.074)	0.2 (0.678)	0.7 (0.405)	4.5 (0.010)	8.7 (<0.001)	0.2 (0.845)
SIR	8.3 (0.004)	19.2 (<0.001)	0.1 (0.825)	23.0 (<0.001)	0.1 (0.715)	0.0 (0.918)	0.2 (0.704)	0.1 (0.917)	6.9 (0.001)	0.4 (0.678)
[Mineral N]†	1.7 (0.191)	32.6 (<0.001)	1.8 (0.183)	16.8 (<0.001)	0.0 (0.923)	0.1 (0.705)	2.1 (0.149)	0.1 (0.931)	11.2 (<0.001)	1.0 (0.372)
DON†	8.3 (0.004)	0.0 (0.891)	0.6 (0.427)	4.7 (0.009)	0.0 (0.993)	1.2 (0.285)	0.4 (0.551)	3.0 (0.049)	0.1 (0.940)	0.1 (0.921)
N ratio	3.2 (0.074)	12.1 (<0.001)	0.4 (0.534)	7.0 (0.001)	0.2 (0.647)	0.8 (0.367)	3.1 (0.078)	3.4 (0.037)	4.9 (0.009)	0.2 (0.815)
Decomposition rate	8.2 (0.005)	32.8 (<0.001)	0.9 (0.343)	9.4 (<0.001)	3.5 (0.061)	1.2 (0.285)	0.6 (0.809)	7.9 (<0.001)	6.9 (0.001)	1.5 (0.212)
MANOVA‡	6.2 (<0.001)	18.0 (<0.001)	1.5 (0.166)	10.0 (<0.001)	1.7 (0.119)	0.5 (0.814)	0.9 (0.478)	2.9 (0.001)	5.6 (<0.001)	0.5 (0.898)

F values (with P values in parentheses) from univariate ANOVA and MANOVA are presented for effects of plant functional group removal, island size and two-way interactions, after seven years. Statistically significant values (at $P = 0.05$) are shown in bold.

* Factors are: R, tree root removal; S, shrub removal; M, moss removal; A, island area class. Results are from three-way univariate ANOVAs for aboveground variables and four-way univariate ANOVAs and MANOVA for belowground variables. Results for three- and four-way interaction terms were rarely statistically significant and are therefore not presented. The degrees of freedom for R, M, S and all interactions among them are 1 for univariate ANOVAs and 6 for MANOVA. For A and all interactions involving A, there are 2 degrees of freedom for univariate ANOVAs and 12 for MANOVA. Residual degrees of freedom are 108, 216 and 211 for aboveground ANOVAs, belowground univariate ANOVAs and MANOVA, respectively.

BR, basal respiration; SIR, substrate-induced respiration; DON, dissolved organic nitrogen; N ratio = (mineral N)/(mineral N + DON).

† Analyses done on (log + 1)-transformed data.

‡ F values estimated from Wilk's lambda.

¹Department of Forest Vegetation Ecology, Faculty of Forestry, Swedish University of Agricultural Sciences, S901-83 Umeå, Sweden. ²Landcare Research, PO Box 69, Lincoln 8152, New Zealand.

We used a 'removal experiment' approach, because this is a powerful tool for investigating the effects of local, non-random losses of biotic components and species interactions in natural ecosystems¹⁶. Fourteen plots, each representing a different removal treatment, were established on each island in August 1996 (420 plots in total). The following removal treatments were performed: a full factorial combination of three plant functional group removals (removal of tree roots by root trenching, ericaceous dwarf shrub removal, and moss removal), and a full factorial combination of three ericaceous dwarf shrub species removals (removal of *Vaccinium myrtillus*, *Vaccinium vitis-idaea* and *Empetrum hermaphroditum*) (see Methods). This design allows assessment of the contributions of all possible interactions among functional groups and among major species within a functional group to ecosystem functioning at local spatial scales across the island size gradient. Here we report on measures of aboveground and belowground ecosystem performance seven years after initiation of the study.

Measures of the cover of each shrub species made on each plot after seven years showed that total shrub cover was unaffected by moss removal (Table 1), but was reduced by an average of 9.9% by root trenching; the magnitude of this effect was independent of island size (Table 1). There was a much stronger influence of shrub species removals on total shrub cover (Table 2 and Fig. 1); all single species removals, and all two-way combinations of removals, had negative effects. This means that at the within-island scale there is a distinct positive relationship between plant diversity and biomass. This is indicative of resource partitioning or resource use complementarity among coexisting species^{1,2,13}. However, this partitioning is only partial; for five out of six cases a given species also showed a positive overall response to the removal of another species (Table 2). Across all the islands, *V. myrtillus* was enhanced by 31.1% and 26.0% by the removal of *V. vitis-idaea* and *E. hermaphroditum*, respectively. *V. vitis-idaea* was enhanced by 19.1% by the removal of *V. myrtillus*, but was unaffected by removal of *E. hermaphroditum*. *E. hermaphroditum* was enhanced by 31.9% and 80.6% by the removal of *V. myrtillus* and *V. vitis-idaea*, respectively. This indicates that after seven years, species have at least partially compensated for removal of other species of the same functional group.

In most cases, the magnitude of the effect on one species of removing another species did not differ significantly across island size classes (Table 2), meaning that the relative balance between resource use complementarity and compensatory effects was not context-dependent. In contrast, the effects on total plant cover of removing two of the three species were significantly influenced by island size class (Table 2 and Fig. 1). Specifically, effects of removing *V. myrtillus* on total cover were strongest on medium and large islands, but removal of *E. hermaphroditum* had the strongest effects on small islands. As the biomass of *V. myrtillus* is greatest on large islands, and *E. hermaphroditum* biomass is maximal on small islands, these results emphasize that within any given ecosystem, effects of species loss are greatest when dominant species are involved¹⁷, and that context-dependent effects of species removals across ecosystems can arise whenever different species dominate in different ecosystems. These effects remained significant after the amount of vegetation removed from each plot at the start of the study was accounted for as a covariate (see Supplementary Information), presumably because the remaining species had at least partially compensated for those species that were removed.

Because the aboveground and belowground subsystems interact to regulate community and ecosystem properties¹⁸, we also assessed a suite of functionally relevant belowground properties for each plot. We found that removals of each of two functional groups (tree roots and shrubs) and each of two species (*V. myrtillus* and *V. vitis-idaea*) influenced several of the measured properties (Tables 1, 2 and Fig. 2). Wherever they had significant effects, these four removals each reduced soil microbial respiration, substrate-induced respiration (SIR; a surrogate for microbial biomass) and litter decomposition

rates, and increased the amount of available mineral nitrogen and the ratio of mineral nitrogen to dissolved organic nitrogen (Fig. 2). Collectively, the trees and two *Vaccinium* species account for the majority of net primary productivity (NPP) on the islands¹⁵, and the removal of these components would serve to reduce NPP, in turn inducing bottom-up limitation of the soil microbial biomass, depressing microbial activity, and impairing functions performed by the microflora, such as decomposition¹⁸. Reduced uptake of nitrogen by plants and microbes after these removals would explain

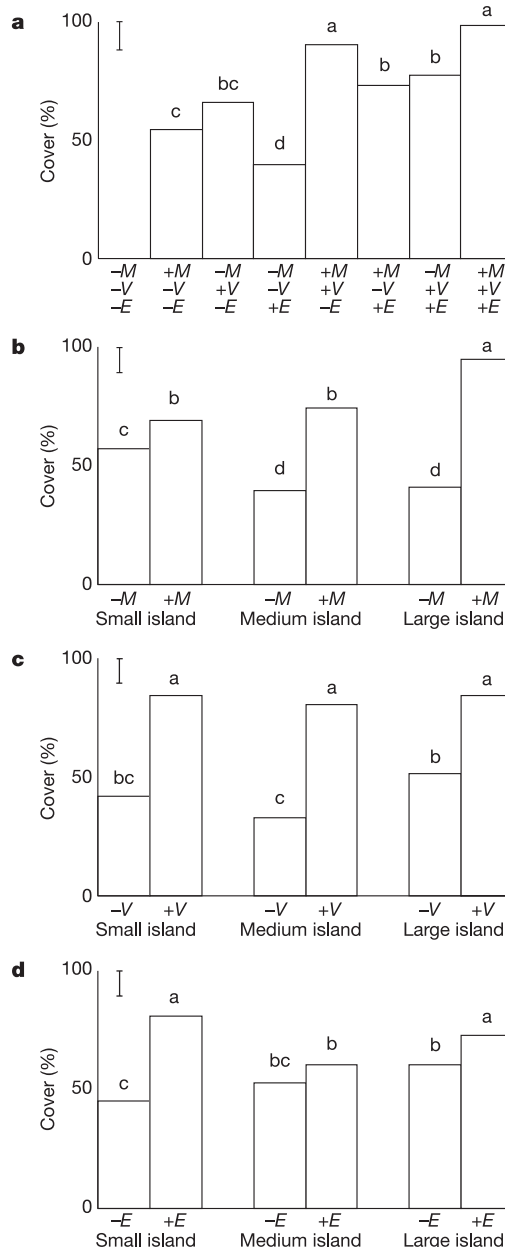


Figure 1 | The influence of shrub species removals on total shrub cover in island ecosystems. **a–d**, Effects of interactions between different shrub species removal treatments (**a**), and between island size and removals of *V. myrtillus* (**b**), *V. vitis-idaea* (**c**) and *E. hermaphroditum* (**d**) on total shrub cover after seven years, measured as total number intercepts per 100 points. Removal treatment codes are: *V. myrtillus* removed (–M) or not removed (+M); *V. vitis-idaea* removed (–V) or not removed (+V); *E. hermaphroditum* removed (–E) or not removed (+E). Within each panel, bars topped by the same letter (a, b, c or d) are not significantly different at $P = 0.05$ (least significant difference (LSD) test), and vertical bars represent LSD values at $P = 0.05$.

the greater amounts of mineral nitrogen present.

Notably, there were significant interactive effects between these removal treatments and island size on belowground properties (Tables 1 and 2), which remained significant when the amount of shrub biomass removed at the start of the study was included as a covariate (see Supplementary Information). Specifically, these removals frequently had significant effects on large and medium sized islands, but never on small ones (Fig. 2). The interactive effects of the two *Vaccinium* species with island size probably emerge because the densities of these species themselves vary across the island size gradient^{11,15}. These species are relatively productive¹⁵, show a high turnover of tissue (see Supplementary Table 1) and produce litter of a relatively high quality¹⁹, which should in turn promote soil biota¹⁹. Therefore, removal of these species should exert greater bottom-up regulation on the soil biota on the islands that they dominate. In contrast, *E. hermaphroditum*, which dominates on small islands, is less productive and produces poorer quality litter that is well defended; loss of this species from islands is therefore less likely to impair decomposer organisms and processes¹⁹. In combination, these results show that belowground effects of losses of biota at either the functional group or species level are context-dependent, and are affected by which species are lost, and the ecosystems that they are lost from.

This study also highlights the importance of spatial scale. It has frequently been suggested that discrepancies across studies on how biodiversity affects ecosystem properties are influenced in part by differences in scale, and that studies at the across-ecosystem scale (representing most observational studies) can yield different results to those at the within-ecosystem scale (representing most experimental studies)^{4,18,20}. Island systems, in which each island operates as

an unambiguously separate ecosystem, have the potential to serve as powerful tests of this idea. Earlier studies on this system have shown that at the across-ecosystem (between island) scale, most ecosystem processes were negatively related to plant diversity, because possible effects of diversity were entirely overridden by plant compositional effects or extrinsic abiotic factors^{5,11,15}. Here we show that at the within-island scale, richness of species in the shrub layer (but not of functional groups) consistently positively affected total shrub cover (and therefore biomass) across the entire island size gradient. Because of the fully factorial nature of the experiment (made possible by the low floristic diversity of the system), these results cannot be explained by statistical artefact, an issue that has created much debate regarding earlier studies^{1,2,18}. These within-island, aboveground effects of plant diversity did not translate consistently belowground: although removal of some biotic components affected soil properties on large islands, this never occurred on small islands. However, whenever removals of species or functional groups influenced key soil processes (respiration or decomposition) within islands, the effects were always negative, indicating the opposite direction of relationship between diversity and function to that observed across islands.

These results have several implications. First, they point to the utility of islands for investigating questions about community and ecosystem ecology. Abiotic and biotic attributes that vary among islands do not just determine the types of communities present^{21–23}, but also how these communities determine ecosystem-level attributes, and in particular what happens to ecosystem functioning when components of island biotas are lost. Second, they show, through the use of independent replicated ecosystems, that components of biodiversity can promote ecosystem process rates within

Table 2 | Effects of shrub species removal, island size and two-way interactions on above- and belowground properties

	Response variable*			
	M	V	E	A
Total shrub cover	119.8 (<0.001)	180.9 (<0.001)	40.5 (<0.001)	4.3 (0.015)
<i>V. myrtillus</i> cover	—	5.9 (0.017)	3.8 (0.049)	37.3 (<0.001)
<i>V. vitis-idaea</i> cover	4.4 (0.039)	—	1.6 (0.205)	6.1 (0.003)
<i>E. hermaphroditum</i> cover	3.7 (0.050)	17.3 (<0.001)	—	41.9 (<0.001)
Moss biomass	0.2 (0.669)	1.6 (0.214)	0.1 (0.708)	1.3 (0.276)
BR	0.9 (0.334)	25.3 (<0.001)	0.2 (0.621)	23.6 (<0.001)
SIR	0.2 (0.664)	17.0 (<0.001)	0.1 (0.739)	12.3 (<0.001)
Mineral N concn†	6.4 (0.012)	10.5 (0.001)	0.8 (0.370)	13.9 (<0.001)
DON†	0.3 (0.581)	0.1 (0.936)	0.2 (0.636)	3.6 (0.028)
N ratio	0.8 (0.359)	5.7 (0.018)	0.2 (0.689)	1.3 (0.271)
Decomposition rate	14.1 (<0.001)	13.8 (<0.001)	0.6 (0.455)	20.2 (<0.001)
MANOVA‡	5.3 (<0.001)	12.0 (<0.001)	0.3 (0.917)	8.4 (<0.001)

	Response variable*					
	M × V	M × E	V × E	M × A	V × A	E × A
Total shrub cover	11.9 (<0.001)	3.9 (0.049)	9.8 (0.002)	15.5 (<0.001)	2.1 (0.129)	7.6 (<0.001)
<i>V. myrtillus</i> cover	—	—	0.4 (0.545)	—	1.0 (0.388)	0.3 (0.712)
<i>V. vitis-idaea</i> cover	—	1.2 (0.285)	—	1.6 (0.218)	—	4.4 (0.014)
<i>E. hermaphroditum</i> cover	0.1 (0.825)	—	—	0.8 (0.471)	0.6 (0.573)	—
Moss biomass	3.9 (0.051)	1.0 (0.323)	1.3 (0.226)	0.7 (0.505)	0.3 (0.709)	0.3 (0.773)
BR	0.2 (0.886)	5.7 (0.018)	0.5 (0.468)	0.7 (0.484)	7.1 (0.001)	0.1 (0.917)
SIR	2.2 (0.139)	0.0 (0.910)	0.1 (0.722)	0.5 (0.615)	5.2 (0.006)	0.2 (0.843)
Mineral N concn†	2.0 (0.163)	0.9 (0.340)	0.6 (0.429)	3.1 (0.045)	4.1 (0.018)	1.0 (0.361)
DON†	0.1 (0.830)	0.1 (0.765)	0.7 (0.657)	0.9 (0.425)	0.0 (0.983)	0.2 (0.835)
N ratio	0.5 (0.482)	0.8 (0.388)	0.1 (0.736)	0.4 (0.645)	1.9 (0.147)	0.6 (0.574)
Decomposition rate	5.1 (0.025)	8.4 (0.004)	3.5 (0.064)	3.7 (0.026)	4.4 (0.014)	0.7 (0.477)
MANOVA‡	1.5 (0.166)	3.1 (0.006)	0.7 (0.646)	1.8 (0.044)	4.4 (<0.001)	0.6 (0.831)

F values (with P values in parentheses) from univariate ANOVA and MANOVA are presented for effects of shrub species removal, island size and two-way interactions on above- and belowground properties after seven years. Statistically significant values (at $P = 0.05$) are shown in bold.

* Factors are: M, *V. myrtillus* removal; V, *V. vitis-idaea* removal; E, *E. hermaphroditum* removal; A, island area class. Results are from three-way univariate ANOVAs for responses of individual species and four-way ANOVA and MANOVA for all other data. The MANOVA incorporates all belowground variables. Results for three- and four-way interaction terms were rarely statistically significant and are not presented. The degrees of freedom for M, V, E and all interactions among them are 1 for univariate ANOVAs and 6 for MANOVA. For A and all interactions involving A, there are 2 degrees of freedom for univariate ANOVAs and 12 for MANOVA. Residual degrees of freedom are 108 for responses of individual species, 216 for univariate ANOVAs on all other response variables, and 211 for MANOVA.

† Analyses done on (log + 1)-transformed data.

‡ F values estimated from Wilk's lambda.

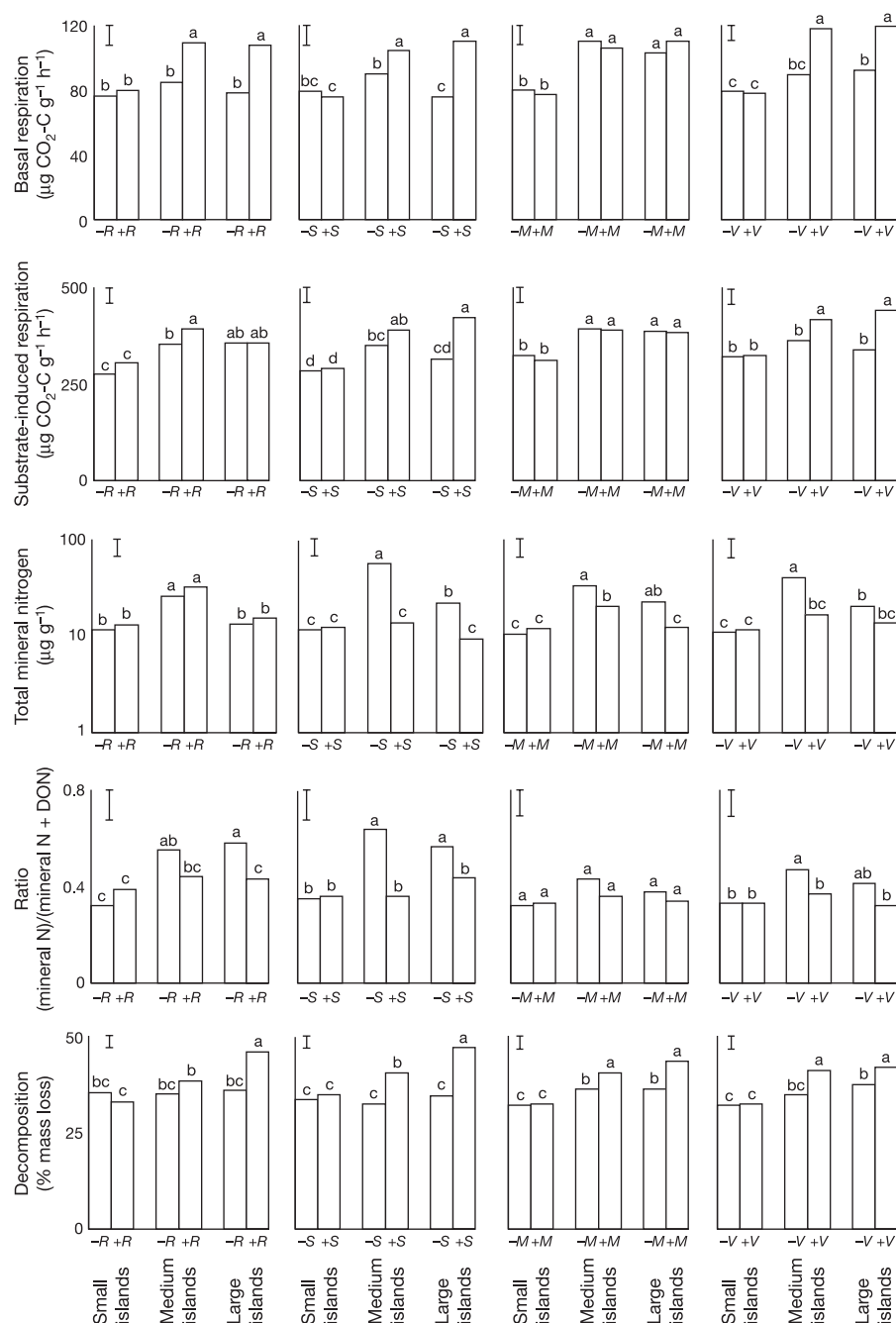


Figure 2 | Interactive effects of island size and removals of functional groups or species on selected belowground properties after seven years. Removal treatments shown are: tree roots removed ($-R$) or not removed ($+R$), all shrubs removed ($-S$) or not removed ($+S$), *Vaccinium myrtillus* removed ($-M$) or not removed ($+M$), and *Vaccinium vitis-idaea* removed ($-V$) or not removed ($+V$). Data for removal of mosses or

E. hermaphroditum are not shown as these treatments did not interact with island size to affect the response variables. DON, dissolved organic nitrogen. Within each panel, bars topped by the same letter (a, b or c) are not significantly different at $P = 0.05$ (LSD test), and vertical bars represent LSD values at $P = 0.05$.

ecosystems, even when biodiversity and process rates are negatively correlated across ecosystems. Finally, they show that the effects of losses of subsets of the resident biota depend on which ecosystems are considered; removals of particular functional groups or species were found to have important effects on some islands but not others, depending on island size and therefore historical disturbance regime and successional stage. This provides evidence that the effects of species and functional groups on ecosystem properties are highly context-dependent, and indicates that seemingly idiosyncratic biodiversity effects can be better understood through comparisons of

independent ecosystems that differ in fundamental abiotic and biotic properties.

METHODS

Plot and treatment setup. We established 14 experimental plots on each of the 30 islands, each representing a different removal treatment of functional groups or species. The experiment was conducted as two components. For the 'functional group removal' component, this consisted of a full factorial combination of three factors (eight treatments in total): tree root removal, ericaceous dwarf shrub removal and moss removal. For the 'species removal' component, this also

consisted of a full factorial combination of three factors (eight treatments in total): removal of *Vaccinium myrtillus*, removal of *Vaccinium vitis-idaea* and removal of *Empetrum hermaphroditum*. These shrubs dominate the ericaceous shrub layer in large, medium and small islands, respectively^{11,15}. Two treatment plots (no removals and removal of all shrubs) were common to both components of the study, yielding 14 plots per island. The three manipulated functional groups represent >99% of all plant biomass present, and the three manipulated shrub species collectively represent 98% of total biomass in the dwarf shrub layer. All plots were 55 cm × 55 cm, but only the inner 45 cm × 45 cm were ever measured or sampled. All plots were located at similar distances from the shore for each island, regardless of island size, to prevent edge and microclimatic effects from confounding the results^{11,15}. These plots are in the vicinity of plots used for other studies previously performed on these islands¹⁵.

The experiment was established in August 1996 (when each treatment was implemented for the first time) and has been maintained annually ever since. Tree root removals have been performed by annual root trenching to below the tree rooting zone²⁴, and dwarf shrub and moss manipulation treatments have been conducted through annual physical removal of vegetation¹⁶. It is recognized that both root trenching and vegetation removals impose initial disturbance effects, but these are likely to be transient^{16,24} and of minimal importance by year seven. Notably, the relative effects of different treatments on vegetation cover performed in these plots varied little after year four. Removal of tree effects was only performed belowground. However, the proportion of total incoming light intercepted by trees (measured on each island in August 2001)¹⁵ is low and therefore unlikely to be sufficient to impair understory vegetation.

Measurements. Every August from 1996 until 2003, the total cover of each ericaceous dwarf shrub species was assessed in each plot by point quadrant analysis, by determining the total number of times the vegetation of that species was intercepted by a total of 100 downwardly projecting points¹⁵. For each of the three species, the total number of point intercepts is very closely correlated with aboveground standing biomass, with R^2 values consistently above 0.90 (ref. 15).

In August 2001, two nylon mesh litterbags (hole size 1.0 × 0.1 mm) were placed in each plot, and buried under about 3 cm of humus. Each litterbag contained 1.5 g of a standardised substrate: locally collected *Salix caprea* leaf litter (0.56% nitrogen, 21% lignin). Litterbags were left in the plots for two years and harvested in August 2003. At that time, dry weight was determined by drying at 80 °C for 24 h.

Soil samples were collected in August 2003. For each plot, three separate soil samples were collected to 5-cm humus depth, bulked within the plot, and sieved to 4 mm. Total nitrogen, phosphorous, ammonium, nitrate and dissolved organic nitrogen (DON) concentrations were measured using automated colourimetric procedures²⁵. Soil basal respiration and substrate-induced respiration (a relative measure of active microbial biomass) were assessed *in vitro* through the use of infrared gas analysis^{26,27}.

Data for all response variables for aboveground variables were analysed by univariate analysis of variance (ANOVA). Belowground variables (for which the same ANOVA model could be applied to all response variables) were analysed by multivariate ANOVA (MANOVA) followed by univariate ANOVAs for each variable. Separate analyses were performed for the functional group removal and species removal components of the study. For each analysis, island size class and removals of each of the species or functional groups were fixed effects, and individual islands served as the units of replication. Data was transformed as necessary to satisfy assumptions of ANOVA and MANOVA.

Received 28 February; accepted 5 April 2005.

1. Loreau, M. *et al.* Biodiversity and ecosystem functioning: current knowledge and future challenges. *Science* **294**, 804–808 (2001).
2. Hooper, D. U. *et al.* Effects of biodiversity on ecosystem functioning: a consensus of current knowledge and needs for future research. *Ecol. Monogr.* **75**, 3–35 (2005).
3. Fridley, J. D. Resource availability dominates and alters the relationship between species diversity and ecosystem productivity in experimental plant communities. *Oecologia* **132**, 271–277 (2002).
4. Hector, A. *et al.* Plant diversity and productivity experiments in European grasslands. *Science* **286**, 1123–1126 (1999).
5. Wardle, D. A. Islands as model systems for understanding how species affect

ecosystem properties. *J. Biogeogr.* **29**, 583–592 (2002).

6. Vitousek, P. M. Oceanic islands as model systems for ecological studies. *J. Biogeogr.* **29**, 573–582 (2002).
7. Anderson, W. B. & Polis, G. A. Nutrient fluxes from water to land: seabirds affect plant nutrient status on Gulf of California islands. *Oecologia* **118**, 324–332 (1999).
8. Naeem, S., Thompson, L. J., Lawler, S. P., Lawton, J. H. & Woodfin, R. M. Declining biodiversity can alter the performance of ecosystems. *Nature* **368**, 734–737 (1994).
9. Tilman, D. *et al.* Diversity and productivity in a long-term grassland experiment. *Science* **294**, 843–845 (2001).
10. Hooper, D. U. & Vitousek, P. M. The effects of plant composition and diversity on ecosystem processes. *Science* **277**, 1302–1305 (1997).
11. Wardle, D. A., Zackrisson, O., Hörnberg, G. & Gallet, C. Influence of island area on ecosystem properties. *Science* **277**, 1296–1299 (1997).
12. Enquist, B. J. & Niklaus, K. J. Invariant scaling relations across tree-dominated communities. *Nature* **410**, 655–660 (2001).
13. Hooper, D. U. & Dukes, D. S. Overyielding among plant functional groups in a long-term experiment. *Ecol. Lett.* **7**, 95–105 (2004).
14. Mulder, C. P. H., Koricheva, J., Huss-Danell, K., Högnberg, P. & Joshi, J. Insects affect relationships between plant species richness and ecosystem processes. *Ecol. Lett.* **2**, 237–246 (1999).
15. Wardle, D. A., Hörnberg, G., Zackrisson, O., Kalela-Brundin, M. & Coomes, D. A. Long term effects of wildfire on ecosystem properties across an island area gradient. *Science* **300**, 972–975 (2003).
16. Diaz, S., Chapin, F. S. III, Symstad, A., Wardle, D. A. & Huenneke, L. Functional diversity revealed through removal experiments. *Trends Ecol. Evol.* **18**, 140–146 (2003).
17. Grime, J. P. Benefits of plant diversity to ecosystems: immediate, filter and founder effects. *J. Ecol.* **86**, 902–910 (1998).
18. Wardle, D. A. *Communities and Ecosystems: Linking the Aboveground and Belowground Components* (Princeton Univ. Press, Princeton, 2002).
19. Wardle, D. A., Nilsson, M.-C., Zackrisson, O. & Gallet, C. An ecosystem level perspective of allelopathy. *Biol. Rev.* **73**, 305–319 (1998).
20. Schmid, B. The species richness–productivity controversy. *Trends Ecol. Evol.* **17**, 113–114 (2002).
21. Gonzalez, A. & Chaneton, E. J. Heterotroph species extinction, abundance and biomass dynamics in an experimentally fragmented microecosystem. *J. Anim. Ecol.* **71**, 594–602 (2002).
22. Spiller, D. A. & Schoener, T. W. Folivory on islands with and without insectivorous lizards: an eight-year study. *Oikos* **78**, 15–22 (1997).
23. Tscharntke, T., Steffan-Deweter, I., Krüess, A. & Thies, C. The contribution of small habitat fragments to the conservation of insect communities of grassland-cropland landscape mosaics. *Ecol. Appl.* **12**, 354–363 (2002).
24. Coomes, D. A. & Grubb, P. J. Impacts of root competition in forests and woodlands: a theoretical framework and review of experiments. *Ecol. Monogr.* **70**, 171–207 (2000).
25. Technicon Instruments. *Individual/Simultaneous Determination of Nitrogen and/or Phosphorus in BD Acid Digests* (Technicon Industrial Systems, Tarrytown, 1997).
26. Anderson, J. P. E. & Domsch, K. H. A physiological method for the quantitative measurement of microbial biomass in soils. *Soil Biol. Biochem.* **10**, 215–221 (1978).
27. Wardle, D. A. Changes in the microbial biomass and metabolic quotient during leaf litter succession in some New Zealand forest and scrubland ecosystems. *Funct. Ecol.* **7**, 346–355 (1993).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Sundberg, P. Bellingham, K. Svarvisdottir, C. Maley, N. Anthes, K. Altgott, N. Rodda and K. Orwin for field assistance at various stages of this work; H. Quirk, M. Karlsson and K. Orwin for laboratory assistance; K. Johansson for help with data management; D. Peltzer for assistance with statistics; and T. H. De Luca, T. Fukami, M.-C. Nilsson and D. Peltzer for comments on earlier versions of this manuscript. This work was funded by the Swedish Natural Sciences Research Council (NFR) and Vetenskapsrådet (VR).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.A.W. (david.wardle@svekl.lu.se).

Average remaining lifetimes can increase as human populations age

Warren C. Sanderson^{1,2} & Sergei Scherbov^{2,3}

Increases in median ages, the most commonly used measure of population ageing^{1,2}, are rapid in today's wealthier countries^{2,3}, and population ageing is widely considered to be a significant challenge to the well-being of citizens there⁴. Conventional measures of age count years since birth; however, as lives lengthen, we need to think of age also in terms of years left until death or in proportion to the expanding lifespan. Here we propose a new measure of ageing: the median age of the population standardized for expected remaining years of life. We show, using historical data and forecasts for Germany, Japan and the United States, that although these populations will be growing older, as measured by their median ages, they will probably experience periods in which they grow younger, as measured by their standardized median ages. Furthermore, we provide forecasts for these countries of the old-age dependency ratio rescaled for increases in life expectancy at birth⁵. These ratios are forecasted to change much less than their unscaled counterparts, and also exhibit periods when the population is effectively growing younger.

Population ageing differs from the ageing of an individual. People who survive grow older with each year they live. Populations, on the other hand, can grow younger. Because a wide variety of matters such as the cost of medical care^{6–8}, retirement⁹, bequests¹⁰, consumption¹¹ and the accumulation of human¹² and tangible^{13,14} capital depend not only on age but also on time left to live, our understanding of population ageing must also reflect both of these factors. Because conventionally measured old-age dependency ratios (the ratio of the number of people at the retirement age and above divided by the number of people in the working ages) have caused worry about the sustainability of pensions¹⁵, it is important to recognize that these ratios, rescaled for life expectancy increases, are forecasted to change comparatively little over the century, suggesting caution in our assessment of long-term pension problems.

Figure 1a–c and Supplementary Table 1 provide information about the unstandardized median age of the population, the standardized median age of the population, using the country's (Germany, Japan, United States) life table in 2000 for standardization, and the remaining life expectancy at the unstandardized median age. All figures pertain to both sexes combined and are calculated using period life tables. There are two types of data: the values through to 2000 are observed, whereas those for future years are based on 1,000 stochastic forecasts.

The median age is the age that divides a population into two numerically equal groups, with half of the people being younger than this age and half older. Life expectancy at the median age is the expected number of years to be lived by a person at the median age. It is also the median remaining life expectancy in the population, with half of the people being at ages with lower remaining life expectancies and half at ages with higher ones. Life expectancy at the median age is especially easy to use as an indicator of ageing because it is comparable both across countries and over time.

Medical care expenditures provide an example where calculating the median remaining life expectancy in a population is useful. Health care costs rise rapidly in the last years of a person's life. The change in the median remaining life expectancy between years is equal to the change in the median time to the onset of that phase of rapidly rising costs.

For many of the decades both the median age and the life expectancy at the median age increase. For the three countries, mortality rates at young ages are now quite low and most of the rise in life expectancies at birth derives from life expectancy increases at the older ages. If the median age of the population remained fixed, remaining life expectancy at the median age would surely increase. However, the essence of population ageing is the increase in median ages. If median ages increase slowly, remaining life expectancies at the median age will increase. On the other hand, median ages can increase so rapidly relative to improvements in mortality rates that remaining life expectancies fall.

An example of a rapid increase in median age outrunning survival rate increases can be seen for Japan between 2000 and 2040 (Fig. 1b). Here, the median age is expected to rise from 41.3 yr to 55.0 yr while the life expectancy at the median age falls from 41.1 yr to 35.0 yr. In the remaining 60 yr of the century, Japan provides an example of where slower increases in the median age are associated with gains in life expectancy at the median age. One broad conclusion from Fig. 1a–c for all three countries is that even in the presence of significant ageing, as measured by increases in the median age, life expectancies at the median age are likely to change only moderately.

Median ages in a country change because of prior changes in fertility, mortality and migration rates. In Japan, the median age is rising rapidly because of a combination of relatively low fertility, high life expectancy and little migration. The United States stands at the opposite end of the spectrum. Its slow increase in median age is a result of relatively high fertility, somewhat lower life expectancy and substantial migration. Germany has demographic rates between those of Japan and the United States.

One disadvantage of using life expectancy at the median age as a measure of ageing is that it is not directly comparable to the median age itself. For comparability it is useful to have another median age, one based on the expected number of years a person has left to live. This is the standardized median age.

The life expectancy standardized population is the hypothetical population that arises when the age of each individual in a specific year is changed to the age of the person in 2000 who had the same remaining life expectancy. For example, if a 40-yr-old person in 2050 had a remaining life expectancy of 50 yr, and a 30 yr old had the same remaining life expectancy (50 yr) in 2000, then the 40-yr-old person would be assigned an age of 30 in the life expectancy standardized population. By definition, when the standardization is done using the country's own life table, the median age and the standardized median age of the population are the same in 2000.

¹Departments of Economics and History, State University of New York at Stony Brook, Stony Brook, New York 11794-4384, USA. ²World Population Project, International Institute for Applied Systems Analysis (IIASA), Laxenburg A-2361, Austria. ³Vienna Institute of Demography, Prinz Eugen Strasse 8, Vienna A-1040, Austria.

Median ages in the three countries generally increase over time; however, standardized median ages show a different pattern of change. In the United States, the standardized median values of the forecasted distributions fall continuously from 2000 onwards, whereas in Germany and Japan, they first increase at the beginning of the century and then decrease. A decreasing standardized median age is far from a certainty in the United States. The 95% prediction intervals for all years from 2010 to 2100 include the value of the standardized median age in 2000 (see Supplementary Table 1). We also show in Supplementary Table 1 that an increase in the standardized median age in first decades of the century seems almost certain in Germany and Japan.

Although we are confident that ageing will occur throughout the century in all three countries as measured by the unstandardized median age, we are also sure that much less ageing or even some increase in youthfulness will be observed using the concept of the standardized median age. When considered from different perspectives, populations in some periods will be growing simultaneously younger and older.

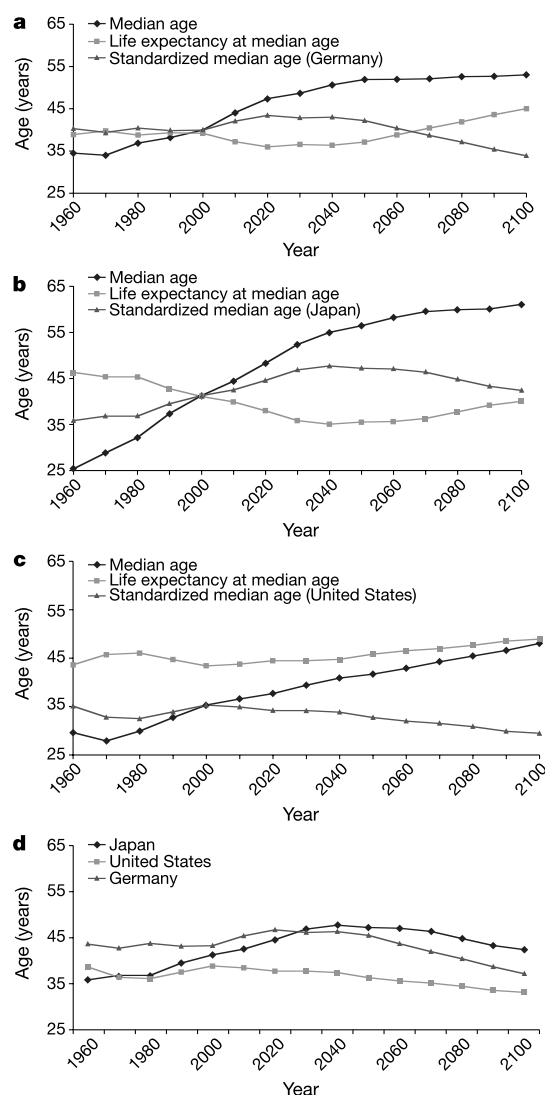


Figure 1 | Unstandardized and standardized median ages, and life expectancies at unstandardized median ages. **a**, Germany; **b**, Japan; **c**, United States. **a–c**, Standardized median ages based on country-specific life tables for 2000. **d**, Standardized median ages based on Japanese life table for 2000. The values through to 2000 are observed; later values are medians based on 1,000 simulations (for 95% prediction intervals see Supplementary Table 1). All values are based on period life tables.

In Fig. 1d, we plot the median ages for the three countries standardized using the 2000 Japanese life table. When the standardization is done with a single country's life table, standardized median ages are comparable across countries. We use Japan as the standard because it had the highest life expectancy among all the countries of the world in 2000. At the beginning of this century, the differences in standardized median ages across the countries were relatively small. The difference between the highest standardized median age (43.2 in Germany) and the lowest (38.8 in the USA) was 4.4 yr. At mid-century the gap in the median forecasts widens significantly to 10.9 yr, with Japan having the highest value and the USA continuing to have the lowest. In 2050, Japan's population will be considerably older than that of the USA both in terms of the unstandardized and standardized median ages.

Figure 2a–c provides a second perspective on ageing using the concept of proportional life cycle rescaling⁵. Proportional life cycle rescaling is a heuristic not a predictive concept. It provides one simple way of thinking about a complex future in which the lengths of life cycle phases will be influenced by social policies and demographic constraints not modelled here. We use proportional life cycle

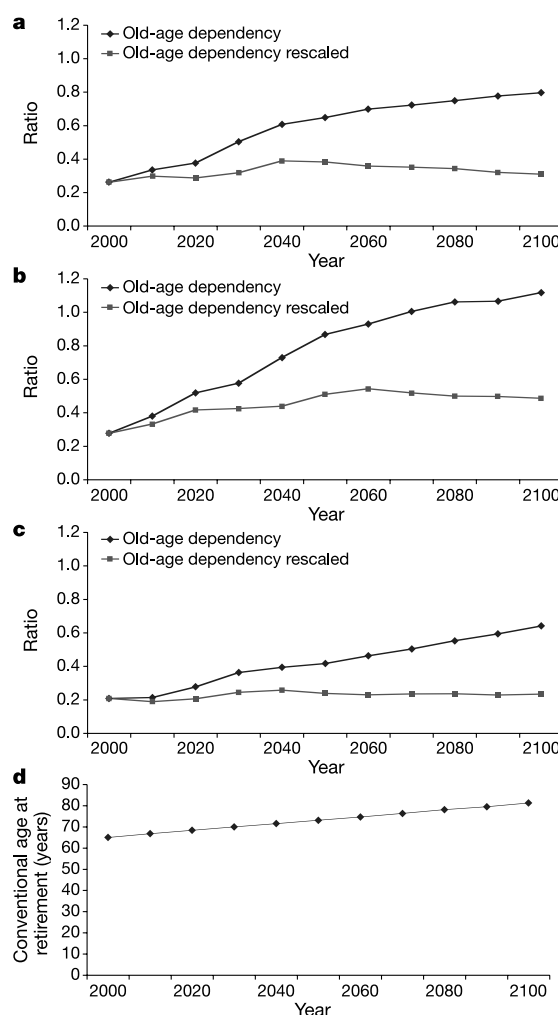


Figure 2 | Conventional and rescaled old-age dependency ratios. **a**, Germany; **b**, Japan; **c**, United States. The old-age dependency ratio is the ratio of the number of people at the retirement age and above divided by the number of people in the working ages. Rescaling increases ages at the beginning and end of working interval proportionally to changes in life expectancy at birth. **d**, Proportionally adjusted retirement ages in Japan. The values for 2000 are observed; later values are medians based on 1,000 simulations (for 95% prediction intervals see Supplementary Table 2). All figures are based on period life tables.

rescaling by adjusting the conventional start of the working age phase (assumed to be age 20 in the year 2000) and the conventional end of that phase (assumed to be age 65 in 2000) proportionally to changes in life expectancy from 2000 onward. Figure 2a–c contains conventional measures of the old-age dependency ratio and new versions of these measures calculated assuming proportional life cycle rescaling.

For all three countries the conventional old-age dependency ratio increases markedly over the century. In Germany, it rises from 0.261 in 2000 to a median forecasted value of 0.797 in 2100. In Japan, the increase is larger, going from 0.276 in 2000 to 1.118 in 2100. In the United States, the increase is smaller than in Germany, but the conventional measure still triples over the century.

The rescaled values show a different pattern. The rescaled old-age dependency ratios rise initially in all three countries and then fall. The rise is quite likely, with the 95% prediction interval in 2040 lying entirely above the ratio in 2000 for all three countries (see Supplementary Table 2). After the middle of the century, changes in the ratio are unclear because the magnitudes of the declines are small relative to the uncertainty involved. For all three countries, the rescaled old-age dependency ratios show considerably less change than the conventional ones.

The new measures presented here are not meant to supplant existing measures, but to supplement them. A perspective that incorporates the new measures presented here is crucial if we are to understand and react appropriately to the challenges of population ageing.

METHODS

The probabilistic forecasts make use of our previously published methodology^{16,17} specialized to the individual countries. In our previous work, we used a mean total fertility rate (TFR) of 2.0 for North America in 2082. Here we use a mean TFR of 1.85 for the United States in that year. This is slightly lower than the one assumed by the United Nations. A lower TFR increases the standardized median age of the population because it results in a smaller number of young people. In our earlier work, we assumed a mean TFR of 1.6 in the region comprised of Japan, Australia and New Zealand in 2082, which is dominated by the population of Japan. We used the same TFR for Japan at that date. We also assumed a mean TFR of 1.7 for Western Europe in 2082. Here we assume a mean TFR of 1.6 for Germany in 2082, because its fertility has been below the average for Western Europe for the last three decades. We assume that distributions around the means are normal with a 90% chance of observing an outcome within half a child of the mean.

Our mortality assumptions are also very similar to those made for the corresponding regions in our earlier work. Life expectancy increases were assumed to have a mean value of 2 yr per decade with a 90% chance of an outcome within 1 yr of the mean. This is consistent with observations over the past four decades¹⁸ and other recently published work¹⁹. Our migration assumptions were made using the same procedure as in our earlier work, except that they were based on observations for the specific countries.

Figure 2d shows the evolution of the rescaled conventional age at retirement in Japan. The paths for Germany and the USA are almost identical. By construction, this age is 65 in 2000. Using our life expectancy forecasts and the proportionality hypothesis, the median forecasted conventional age rises to 73 by 2050 and continues to climb for the remainder of the century. These rescaled conventional ages are used in the production of Fig. 2.

Forecasted data in Figs 1 and 2 are the median values of the forecast distributions based on 1,000 simulations. The median values and their 95% prediction intervals are presented in Supplementary Tables 1 and 2. In each year, we compute the median age of the population. These median ages depend on the age distribution of the population at the beginning of the forecast period and on the whole time paths of fertility, mortality and migration rates from the beginning of the forecast period to the year in question.

We calculate the distribution of remaining life expectancy at the median age in year t , for example, using the stochastic life table associated with that year. Thus, life expectancies at the median age in year t have uncertainty due to variability in median ages in year t and due to the randomness in the life tables for that particular year. The life table used in period t is closely associated with the time path of life tables before year t and therefore with the age structure and the median age of the population in that year. Standardized median ages are also

subject to both sorts of uncertainty.

The uncertainty in the distributions of remaining life expectancy and the distributions of standardized ages are influenced by the correlations between the median age at time t and life expectancy at time t . The autocorrelation of life expectancies implies that a high life expectancy at time t is associated, on average, with a high median age in that year. High life expectancies and high median ages have opposing effects on the remaining life expectancy at the median age, reducing the uncertainty relative to what it would have been in the absence of those correlations.

The differences in sources of uncertainty can be seen in Supplementary Table 1 by comparing the size of the 95% prediction intervals for unstandardized and standardized median ages. For Germany in 2020, for example, the difference between the upper and lower bounds of the 95% prediction interval for the unstandardized median age is 1.8 yr. The comparable difference for the standardized median age (using either the German or Japanese standards) is 3.8 yr.

The observed median ages and the life expectancies at the median age move less regularly than the forecasted medians. The observed figures take baby booms and busts into account differently. The observed figures are from a single random path of realizations for each country. The forecasted medians essentially average across possible future paths and are therefore much smoother.

Received 13 January; accepted 24 March 2005.

- Gavrilov, L. A. & Heuveline, P. in *The Encyclopedia of Population* Vol. 1 (eds Demeny, P. & McNicoll, G.) 32–37 (Macmillan Reference USA, New York, 2003).
- United Nations. *World Population Prospects—The 2004 Revision: Highlights* (http://www.un.org/esa/population/publications/WPP2004/2004Highlights_finalrevised.pdf) (2005).
- Lutz, W., Sanderson, W., Scherbov, S. & O'Neill, B. in *The End of World Population Growth in the 21st Century: New Challenges for Human Capital Formation and Sustainable Development* (eds Lutz, W., Sanderson, W. & Scherbov, S.) 85–120 (Earthscan, London, 2004).
- Peterson, P. *Gray Dawn* (Random House, New York, 1999).
- Lee, R. D. & Goldstein, J. R. in *Life Span: Evolutionary, Ecological, and Demographic Perspectives. A Supplement to Population and Development Review* Vol. 29 (eds Carey, J. R. & Tuljapourkar, S.) 183–207 (Population Council, New York, 2003).
- Miller, T. Increasing longevity and Medicare expenditures. *Demography* **38**, 215–226 (2001).
- Seshamani, M. & Gray, A. M. A longitudinal study of the effects of age and time to death on hospital costs. *J. Health Econ.* **23**, 217–235 (2004).
- Stearns, S. C. & Norton, E. C. Time to include time to death? The future of health care expenditure predictions. *Health Econ.* **13**, 315–327 (2004).
- Hurd, M., Smith, J. P. & Zissimopoulos, J. The effects of subjective survival on retirement and social security claiming. *J. Appl. Econometr.* **19**, 761–775 (2004).
- Gan, L., Gong, G., Hurd, M. & McFadden, D. *Subjective Mortality Risks and Bequests* (NBER Working Paper 10789, National Bureau of Economic Research, Cambridge, Massachusetts, 2004).
- Lee, R. & Tuljapourkar, S. Death and taxes: Longer life, consumption, and social security. *Demography* **34**, 67–81 (1997).
- Creighton, S. & Hudson, A. L. *Participation Patterns in Adult Education, 1991–1999* (US Department of Education, Washington DC, 2002).
- Bloom, D. E., Canning, D. & Graham, B. Longevity and life-cycle savings. *Scand. J. Econ.* **105**, 319–338 (2003).
- Higgins, M. Demography, national savings, and international capital flows. *Int. Econ. Rev.* **39**, 343–369 (1998).
- Disney, R. Crisis in public pension programmes in OECD: What are the reform options? *Econ. J.* **110**, F1–F23 (2000).
- Lutz, W., Sanderson, W. & Scherbov, S. The end of world population growth. *Nature* **412**, 543–545 (2001).
- Lutz, W., Sanderson, W. & Scherbov, S. in *The End of World Population Growth in the 21st Century: New Challenges for Human Capital Formation and Sustainable Development* (eds Lutz, W., Sanderson, W. & Scherbov, S.) 17–84 (Earthscan, London, 2004).
- Sanderson, W. & Scherbov, S. *Putting Oeppen and Vaupel to Work: On the Road to New Stochastic Mortality Forecasts* (Interim Report IR-04-049, International Institute for Applied Systems Analysis, Laxenburg, Austria, 2004).
- Oeppen, J. & Vaupel, J. Broken limits to life expectancy. *Science* **296**, 1029–1031 (2002).

Supplementary Information is linked to the online version of this paper at www.nature.com/nature.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to W.C.S. (wsanderson@notes.cc.sunysb.edu).

LETTERS

Uncovering the overlapping community structure of complex networks in nature and society

Gergely Palla^{1,2}, Imre Derényi², Illés Farkas¹ & Tamás Vicsek^{1,2}

Many complex systems in nature and society can be described in terms of networks capturing the intricate web of connections among the units they are made of^{1–4}. A key question is how to interpret the global organization of such networks as the co-existence of their structural subunits (communities) associated with more highly interconnected parts. Identifying these *a priori* unknown building blocks (such as functionally related proteins^{5,6}, industrial sectors⁷ and groups of people^{8,9}) is crucial to the understanding of the structural and functional properties of networks. The existing deterministic methods used for large networks find separated communities, whereas most of the actual networks are made of highly overlapping cohesive groups of nodes. Here we introduce an approach to analysing the main statistical features of the interwoven sets of overlapping communities that makes a step towards uncovering the modular structure of complex systems. After defining a set of new characteristic quantities for the statistics of communities, we apply an efficient technique for exploring overlapping communities on a large scale. We find that overlaps are significant, and the distributions we introduce reveal universal features of networks. Our studies of collaboration, word-association and protein interaction graphs show that the web of communities has non-trivial correlations and specific scaling properties.

Most real networks typically contain parts in which the nodes (units) are more highly connected to each other than to the rest of the network. The sets of such nodes are usually called clusters, communities, cohesive groups or modules^{8,10,11–13}; they have no widely accepted, unique definition. In spite of this ambiguity, the presence of communities in networks is a signature of the hierarchical nature of complex systems^{5,14}. The existing methods for finding communities in large networks are useful if the community structure is such that it can be interpreted in terms of separated sets of communities (see Fig. 1b and refs 10, 15, 16–18). However, most real networks are characterized by well-defined statistics of overlapping and nested communities. This can be illustrated by the numerous communities that each of us belongs to, including those related to our scientific activities or personal life (school, hobby, family) and so on, as shown in Fig. 1a. Furthermore, members of our communities have their own communities, resulting in an extremely complicated web of the communities themselves. This has long been understood by sociologists¹⁹ but has never been studied systematically for large networks. Another, biological, example is that a large fraction of proteins belong to several protein complexes simultaneously²⁰.

In general, each node i of a network can be characterized by a membership number m_i , which is the number of communities that the node belongs to. In turn, any two communities α and β can share $s_{\alpha,\beta}^{ov}$ nodes, which we define as the overlap size between these communities. Naturally, the communities also constitute a network,

with the overlaps being their links. The number of such links of community α can be called its community degree, d_{α}^{com} . Finally, the size s_{α}^{com} of any community α can most naturally be defined as the number of its nodes. To characterize the community structure of a large network we introduce the distributions of these four basic quantities. In particular we focus on their cumulative distribution

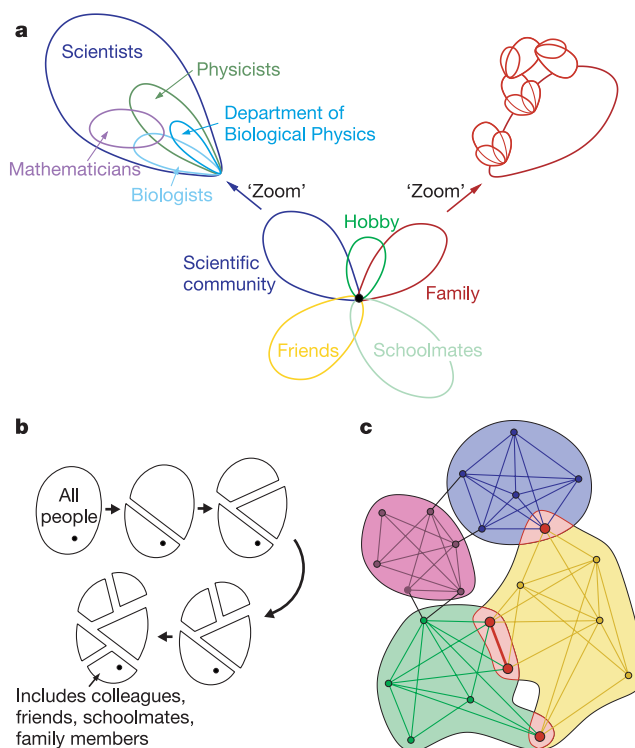


Figure 1 | Illustration of the concept of overlapping communities. **a**, The black dot in the middle represents either of the authors of this paper, with several of his communities around. Zooming in on the scientific community demonstrates the nested and overlapping structure of the communities, and depicting the cascades of communities starting from some members exemplifies the interwoven structure of the network of communities.

b, Divisive and agglomerative methods grossly fail to identify the communities when overlaps are significant. **c**, An example of overlapping k -clique communities at $k = 4$. The yellow community overlaps the blue one in a single node, whereas it shares two nodes and a link with the green one. These overlapping regions are emphasized in red. Notice that any k -clique (complete subgraph of size k) can be reached only from the k -cliques of the same community through a series of adjacent k -cliques. Two k -cliques are adjacent if they share $k - 1$ nodes.

¹Biological Physics Research Group of the Hungarian Academy of Sciences, Pázmány P. stny. 1A, H-1117 Budapest, Hungary. ²Department of Biological Physics, Eötvös University, Pázmány P. stny. 1A, H-1117 Budapest, Hungary.

functions denoted by $P(m)$, $P(s^{ov})$, $P(d^{com})$ and $P(s^{com})$. For the overlap size, for example, $P(s^{ov})$ means the proportion of those overlaps that are larger than s^{ov} . Further relevant statistical features will be introduced later.

The basic observation on which our community definition relies is that a typical community consists of several complete (fully connected) subgraphs that tend to share many of their nodes. Thus, we define a community, or more precisely a k -clique community, as a union of all k -cliques (complete subgraphs of size k) that can be reached from each other through a series of adjacent k -cliques (where adjacency means sharing $k - 1$ nodes)^{21–23}. This definition seeks to represent the fact that it is an essential feature of a community that its members can be reached through well-connected subsets of nodes. There are other parts of the whole network that are not reachable from a particular k -clique, but they potentially contain further k -clique communities. In turn, a single node can belong to several communities. All these can be explored systematically and can result in many overlapping communities (illustrated in Fig. 1c). In most cases, relaxing this definition (for example, by allowing incomplete k -cliques) is practically equivalent to decreasing k . For finding meaningful communities, the way in which they are identified is expected to satisfy several basic requirements: it cannot be too restrictive, it should be based on the density of links, it is required to be local, it should not yield any cut-node or cut-link (whose removal would disjoin the community) and, of course, it should allow overlaps. We employ the community definition specified above, because none of the others in the literature satisfy all these requirements simultaneously^{21,24}.

Although the numerical determination of the full set of k -clique communities is a polynomial problem, we use an algorithm (which can be downloaded from <http://angel.elte.hu/clustering/>) that is exponential, because it is significantly more efficient for the graphs corresponding to real data. This method is based on first locating all cliques (maximal complete subgraphs) of the network and then identifying the communities by carrying out a standard component analysis of the clique–clique overlap matrix²¹. More details about the method and its speed are given in Supplementary Information.

We use our method for binary networks (that is, with undirected and unweighted links). An arbitrary network can always be transformed into a binary one by ignoring any directionality in the links and keeping only those that are stronger than a threshold weight w^* . Changing the threshold is like changing the resolution (as in a microscope) with which the community structure is investigated: by increasing w^* the communities start to shrink and fall apart. A similar effect can be observed by changing the value of k as well: increasing k makes the communities smaller and more disintegrated but also at the same time more cohesive.

When we are interested in the community structure around a particular node, it is advisable to scan through some ranges of k and w^* and monitor how its communities change. As an illustration, in Fig. 2 we show diagrams of the communities of three selected nodes of three large networks: the social network of scientific collaborators²⁵ (Fig. 2a), the network of word associations²⁶ related to cognitive sciences (Fig. 2b) and the molecular-biological network of protein–protein interactions²⁷ (Fig. 2c). These pictures can serve as tests or validations of the efficiency of our algorithm. In particular,

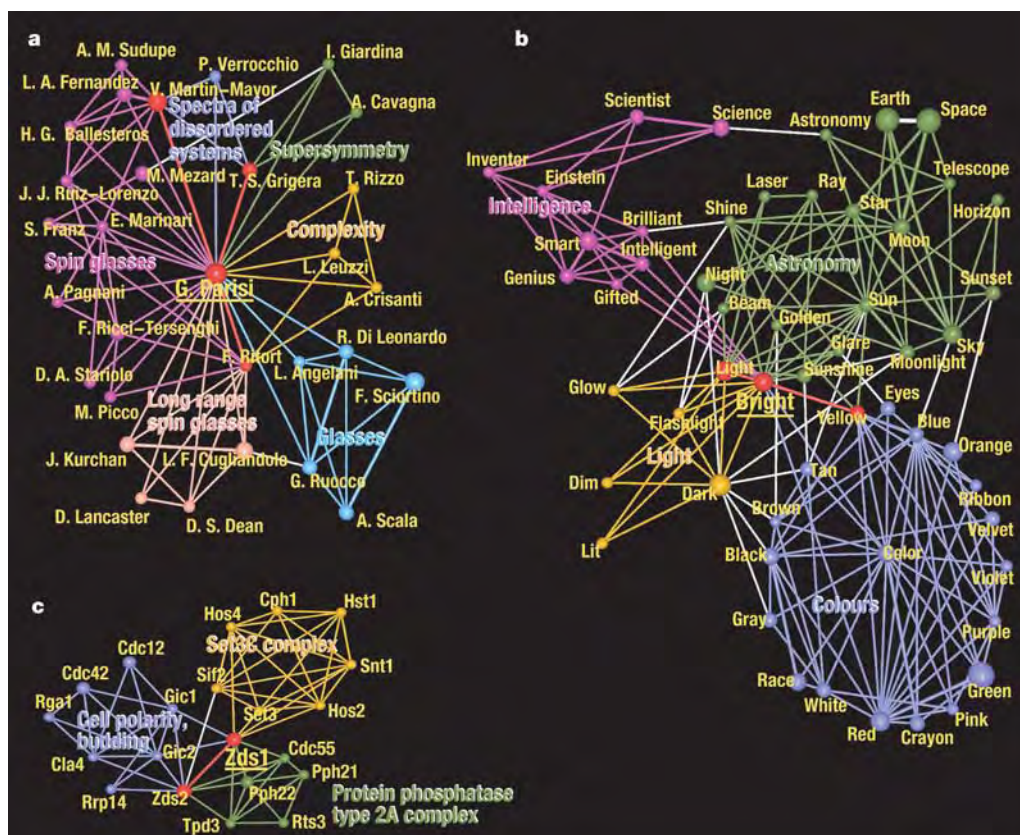


Figure 2 | The community structure around a particular node in three different networks. The communities are colour coded, the overlapping nodes and links between them are emphasized in red, and the volume of the balls and the width of the links are proportional to the total number of communities they belong to. For each network the value of k has been set to 4. **a**, The communities of G. Parisi in the co-authorship network of the Los Alamos Condensed Matter archive (for threshold weight $w^* = 0.75$) can

be associated with his fields of interest. **b**, The communities of the word 'bright' in the South Florida Free Association norms list (for $w^* = 0.025$) represent the different meanings of this word. **c**, The communities of the protein Zds1 in the DIP core list of the protein–protein interactions of *S. cerevisiae* can be associated with either protein complexes or certain functions.

the communities of G. Parisi (whose contributions in different fields of physics are well known) shown in Fig. 2a are associated with his fields of interest, as can be deduced from the titles of the papers involved. The four-clique communities of the word ‘bright’ (Fig. 2b) correspond to the various meanings of this word. An important biological application is finding the communities of proteins, based on their interactions. Indeed, most proteins in the communities shown in Figs 2c and 3 can be associated with either protein complexes or certain functions, as can be looked up by using the GO-TermFinder package²⁸ and the online tools of the *Saccharomyces* Genome Database (SGD)²⁹. For some proteins no function is yet available. Thus, the fact that they show up in our approach as members of communities can be interpreted as a prediction of their functions. One such example can be seen in the enlarged

portion of Fig. 3. For the protein Ycr072c, which is required for the viability of the cell and appears in the dark green community on the right, SGD provides no biological process (function). By far the most significant GO term for the biological process of this community is ‘ribosome biogenesis/assembly’. We can therefore infer that Ycr072c is likely to be involved in this process. In addition, new cellular processes can be predicted if as yet unknown communities are found with our method.

These examples (and further examples included in Supplementary Information) show the advantages of our approach over the existing divisive and agglomerative methods recently used for large real networks. Divisive methods cut the network into smaller and smaller pieces, and each node is forced to remain in only one community and be separated from its other communities, most of which then necessarily fall apart and disappear. This happens, for example, with the word ‘bright’ when we apply the method described in ref. 16: it tends to stay together mostly with the words of the community related to ‘light’, while most of its other communities (for example, those related to ‘colours’; see Fig. 2b) completely disintegrate (‘green’ becomes associated with the vegetables, ‘orange’ with the fruits, and so on). Agglomerative methods do the same, but in the reverse direction. For example, when we applied the agglomerative method of ref. 18, at some point ‘bright’, as a single word, joined a ‘community’ of 890 other words. In addition, such methods inevitably lead to a tree-like hierarchical rendering of the communities, whereas our approach allows the construction of an unconstrained network of communities.

The networks chosen above have been constructed in the following ways. In the co-authorship network of the Los Alamos e-print archives²⁵ each article contributes a value $1/(n-1)$ to the weight of the link between every pair of its n authors. In the South Florida Free Association norms list²⁶ the weight of a directed link from one word to another indicates the frequency with which the people in the survey associated the end point of the link with its starting point. For our purposes these directed links have been replaced by undirected ones with a weight equal to the sum of the weights of the corresponding two oppositely directed links. In the Database of Interacting Proteins (DIP) core list of the protein–protein interactions of *Saccharomyces cerevisiae*²⁷ each interaction represents an unweighted link between the interacting proteins. These networks are very large, consisting of 30,739, 10,617 and 2,609 nodes and 136,065, 63,788 and 6,355 links, respectively.

Although different values of k and w^* might be optimal for the local community structure around different nodes, we should set some global criterion to fix their values if we wish to analyse the statistical properties of the community structure of the entire network. The criterion we use is based on finding a community structure that is as highly structured as possible. In the related percolation phenomena²³ a giant component appears when the number of links is increased above some critical point. Therefore, to approach this critical point from below, for each selected value of k (typically between 3 and 6) we lower the threshold w^* until the largest community becomes twice as big as the second largest one. In this way we ensure that we find as many communities as possible, without the negative effect of having a giant community that would smear out the details of the community structure by merging many smaller communities. We denote by f^* the fraction of links stronger than w^* ,

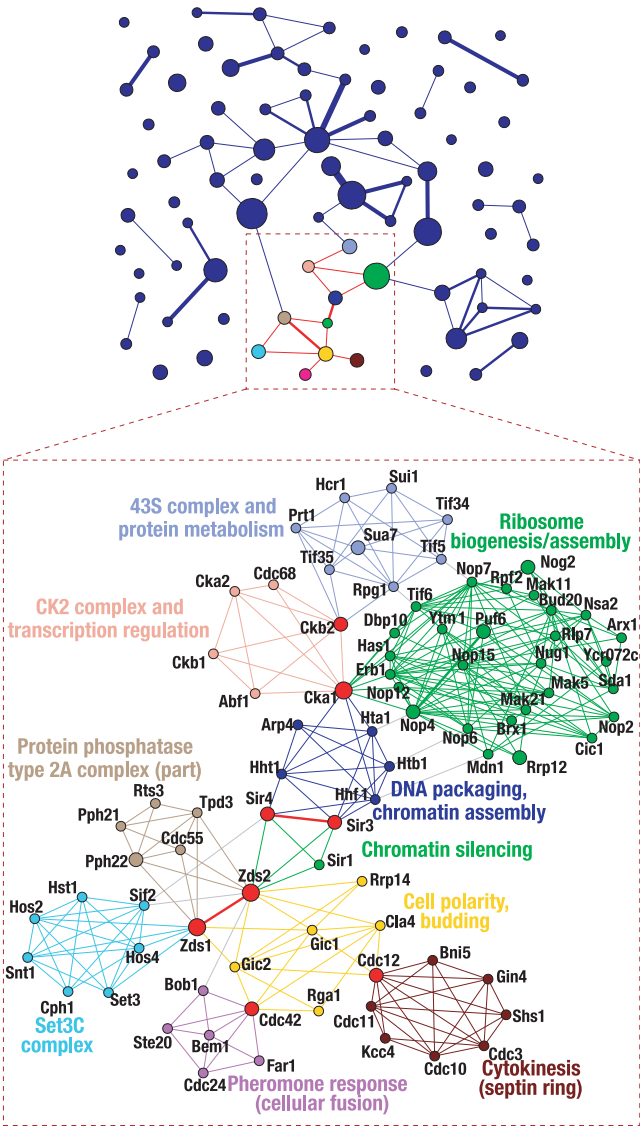


Figure 3 | Network of the 82 communities in the DIP core list of the protein–protein interactions of *S. cerevisiae* for $k = 4$. The areas of the circles and the widths of the links are proportional to the size of the corresponding communities (s_{α}^{com}) and to the size of the overlaps ($s_{\alpha\beta}^{ov}$), respectively. The coloured communities (top) are cut out and magnified to reveal their internal structure (bottom): the nodes and links of the original network have the same colour as their communities, those that are shared by more than one community are emphasized in red, and the grey links are not part of these communities. The areas of the circles and the widths of the links are proportional to the total number of communities they belong to.

Table 1 Statistical properties of the network of communities				
Network	N^{com}	$\langle d^{com} \rangle$	$\langle C^{com} \rangle$	$\langle r \rangle$
Co-authorship	2,450	12.10	0.44	0.58
Word association	670	11.33	0.56	0.72
Protein interaction	82	1.54	0.17	0.26

N^{com} is the number of communities, $\langle d^{com} \rangle$ is the average community degree, $\langle C^{com} \rangle$ is the average clustering coefficient of the network of communities, and $\langle r \rangle$ is the average fraction of shared nodes in the communities.

and use only those values of k for which f^* is not too small (not smaller than 0.5). This has led us to $k = 6$ and $k = 5$ with $f^* = 0.93$ and 0.75, respectively, for the collaboration network, and $k = 4$ with $f^* = 0.67$ for the word-association network. For the former network both sets of parameters result in very similar communities (see Supplementary Information). Because for unweighted networks no threshold weight can be set, for these we simply select the smallest value of k for which no giant community appears. For the protein interaction network this gives $k = 4$, resulting in 82 communities. Because of this relatively low number, we can depict the entire network of protein communities as in Fig. 3.

The four distributions characterizing the global community structure of these networks are shown in Fig. 4. Although the scaling of the size of non-overlapping communities has already been shown

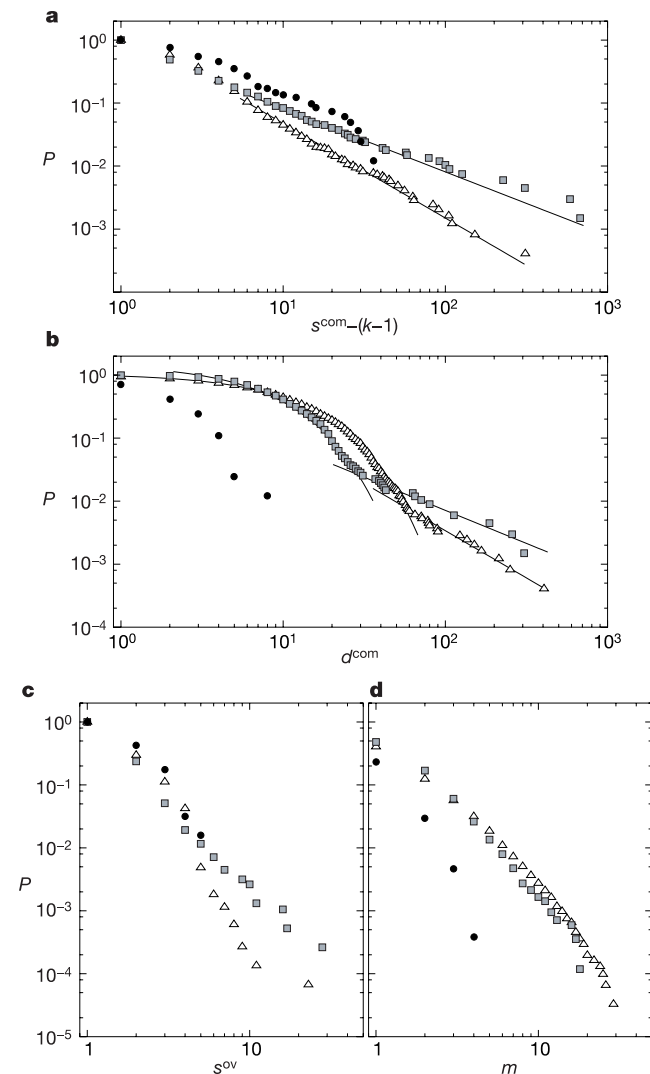


Figure 4 | Statistics of the k -clique communities for three large networks. The networks are the co-authorship network of the Los Alamos Condensed Matter archive (triangles, $k = 6$, $f^* = 0.93$), the word-association network of the South Florida Free Association norms (squares, $k = 4$, $f^* = 0.67$), and the protein interaction network of the yeast *S. cerevisiae* from the DIP database (circles, $k = 4$). **a**, The cumulative distribution function of the community size follows a power law with exponents between -1 (upper line) and -1.6 (lower line). **b**, The cumulative distribution of the community degree starts exponentially and then crosses over to a power law (with the same exponent as for the community size distribution). **c**, The cumulative distribution of the overlap size. **d**, The cumulative distribution of the membership number.

for social networks^{17,18}, it is striking to observe how this aspect of large real networks is preserved even when a more complete picture (allowing overlaps) is investigated. In Fig. 4a the power-law dependence $P(s^{\text{com}}) \propto (s^{\text{com}})^{-\tau}$ with an exponent ranging between $\tau = 1$ and $\tau = 1.6$ is well pronounced and is valid over nearly the entire range of community sizes.

It is well known²⁻⁴ that the nodes of large real networks have a power-law degree distribution. Will the same kind of distribution hold when we move to the next level of organization and consider the degrees of the communities? We find that it is not so. The community degrees (Fig. 4b) have a unique distribution, consisting of two distinct parts: an exponential decay $P(d^{\text{com}}) \propto \exp(-d^{\text{com}}/d_0^{\text{com}})$ with a characteristic community degree d_0^{com} (which is of the order of $\langle d^{\text{com}} \rangle$ shown in Table 1), followed by a power-law tail proportional to $(d^{\text{com}})^{-\tau}$. This new kind of behaviour is consistent with the community size distribution if we assume that, on average, each node of a community has a contribution δ to the community degree. The tail of the community degree distribution is therefore simply proportional to that of the community size distribution. At the first part of $P(d^{\text{com}})$, in contrast, a characteristic scale $d_0^{\text{com}} \approx k\delta$ appears, because most of the communities have a size of the order of k (see Fig. 4a) and their distribution around d_0^{com} dominates this part of the curve. Thus, the degree to which $P(d^{\text{com}})$ deviates from a simple scaling depends on k or, in other words, on the prescribed minimum cohesiveness of the communities.

The extent to which different communities overlap is also a relevant property of a network. Although the range of overlap sizes is limited, the behaviour of the cumulative overlap size distribution $P(s^{\text{ov}})$, shown in Fig. 4c, is close to a power law for each network, with a rather large exponent. We can conclude that there is no characteristic overlap size in the networks. Finally, in Fig. 4d we display the cumulative distribution of the membership number $P(m)$. These plots demonstrate that a node can belong to several communities. In the collaboration and word-association networks there seems to be no characteristic value for the membership number: the data are close to a power-law dependence, with a large exponent. However, in the protein interaction network the largest membership number is only 4, which is consistent with the also rather short distribution of its community degree. To show that the communities we find are not due to an artefact of our method, we have also determined the above distributions for 'randomized' graphs with parameters (size, degree sequence, k and f^*) the same as in our three examples but with links stochastically redistributed between the nodes. We have found that the distributions are indeed extremely truncated, signifying a complete lack of the rich community structure determined for the original data.

In Table 1 we have collected a few statistical properties of the network of communities. It should be pointed out that the average clustering coefficients $\langle C^{\text{com}} \rangle$ are relatively high, indicating that two communities overlapping with a given community are likely to overlap with each other as well, mostly because they all share the same overlapping region. The high fraction of shared nodes is yet another indication of the importance of overlaps between the communities.

The specific scaling of the community degree distribution is a hitherto undescribed signature of the hierarchical nature of the systems we study. We find that if we consider the network of communities instead of the nodes themselves, we still observe a degree distribution with a fat tail, but a characteristic scale appears, below which the distribution is exponential. This is consistent with our understanding of a complex system having different levels of organization with units specific to each level. In the present case the principle of organization (scaling) is preserved (with some specific modifications) when going to the next level, in good agreement with the recent finding of the self-similarity of many complex networks³⁰.

With recent technological advances, huge sets of data are accumulating at a tremendous pace in various fields of human activity

(including telecommunications, the Internet and stock markets) and in many areas of life and social sciences (such as biomolecular assays, genetic maps and groups of World Wide Web users). Understanding both the universal and specific features of the networks associated with these data has become a significant task. The knowledge of the community structure enables the prediction of some essential features of the systems under investigation. For example, because with our approach it is possible to 'zoom' in on a single unit in a network and uncover its communities (and the communities connected to these, and so on), we provide a tool with which to interpret the local organization of large networks and can predict how the modular structure of the network changes if a unit is removed (for example, in a gene knockout experiment). A unique feature of our method is that we can simultaneously look at the network at a higher level of organization and locate the communities that have a key role within the web of communities. Among the many possible applications is a more sophisticated approach to the spreading of infections (for example, real or computer viruses) or information in highly modular complex systems.

Received 17 January; accepted 7 April 2005.

- Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).
- Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
- Albert, R. & Barabási, A.-L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97 (2002).
- Mendes, J. F. F. & Dorogovtsev, S. N. *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford Univ. Press, Oxford, 2003).
- Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. & Barabási, A.-L. Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551–1555 (2002).
- Spirin, V. & Mirny, L. A. Protein complexes and functional modules in molecular networks. *Proc. Natl Acad. Sci. USA* **100**, 12123–12128 (2003).
- Onnela, J.-P., Chakraborti, A., Kaski, K., Kertész, J. & Kanto, A. Dynamics of market correlations: Taxonomy and portfolio analysis. *Phys. Rev. E* **68**, 056110 (2003).
- Scott, J. *Social Network Analysis: A Handbook* 2nd edn (Sage, London, 2000).
- Watts, D. J., Dodds, P. S. & Newman, M. E. J. Identity and search in social networks. *Science* **296**, 1302–1305 (2002).
- Shiffrin, R. M. & Börner, K. Mapping knowledge domains. *Proc. Natl Acad. Sci. USA* **101**, 5183–5185 (2004).
- Everitt, B. S. *Cluster Analysis* 3rd edn (Edward Arnold, London, 1993).
- Knudsen, S. *A Guide to Analysis of DNA Microarray Data* 2nd edn (Wiley-Liss, New York, 2004).
- Newman, M. E. J. Detecting community structure in networks. *Eur. Phys. J. B* **38**, 321–330 (2004).
- Vicsek, T. The bigger picture. *Nature* **418**, 131 (2002).
- Blatt, M., Wiseman, S. & Domany, E. Super-paramagnetic clustering of data. *Phys. Rev. Lett.* **76**, 3251–3254 (1996).
- Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl Acad. Sci. USA* **99**, 7821–7826 (2002).
- Radicchi, F., Castellano, C., Cecconi, F., Loreto, V. & Parisi, D. Defining and identifying communities in networks. *Proc. Natl Acad. Sci. USA* **101**, 2658–2663 (2004).
- Newman, M. E. J. Fast algorithm for detecting community structure in networks. *Phys. Rev. E* **69**, 066133 (2004).
- Faust, K. in *Models and Methods in Social Network Analysis* (eds Carrington, P., Scott, J. & Wasserman, S.) 117–147 (Cambridge Univ. Press, New York, 2005).
- Gavin, A. C. et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
- Everett, M. G. & Borgatti, S. P. Analyzing clique overlap. *Connections* **21**, 49–61 (1998).
- Batagelj, V. & Zaversnik, M. Short cycles connectivity. *arXiv cs.DS/0308011* (<http://arxiv.org/abs/cs/0308011>) (2003).
- Derényi, I., Palla, G. & Vicsek, T. Clique percolation in random networks. *Phys. Rev. Lett.* **94**, 160202 (2005).
- Kosub, S. in *Network Analysis* (eds Brandes, U. & Erlebach, T.) 112–142 (Lecture Notes in Computer Science 3418, Springer, Berlin, 2005).
- Warner, S. E-prints and the Open Archives Initiative. *Library Hi Tech* **21**, 151–158 (2003).
- Nelson, D. L., McEvoy, C. L. & Schreiber, T. A. The University of South Florida word association, rhyme, and word fragment norms. (<http://www.usf.edu/FreeAssociation/>).
- Xenarios, I. et al. DIP: the Database of Interacting Proteins. *Nucleic Acids Res.* **28**, 289–291 (2000).
- Boyle, E. I. et al. GO:TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* **20**, 3710–3715 (2004).
- Cherry, J. M. et al. Genetic and physical maps of *Saccharomyces cerevisiae*. *Nature* **387**, 675–735 (1997).
- Song, C., Havlin, S. & Makse, H. A. Self-similarity of complex networks. *Nature* **433**, 392–395 (2005).

Supplementary Information accompanies the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A.-L. Barabási and P. Pollner for discussions, and B. Kovács and G. Szabó for help with visualization and software support. This research was supported by the Hungarian Research Grant Foundation (OTKA).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to T.V. (vicsek@angel.elte.hu).

Nitrogen transfer in the arbuscular mycorrhizal symbiosis

Manjula Govindarajulu¹, Philip E. Pfeffer², Hairu Jin², Jehad Abubaker¹, David D. Douds², James W. Allen³, Heike Bücking³, Peter J. Lammers¹ & Yair Shachar-Hill³

Most land plants are symbiotic with arbuscular mycorrhizal fungi (AMF), which take up mineral nutrients from the soil and exchange them with plants for photosynthetically fixed carbon. This exchange is a significant factor in global nutrient cycles¹ as well as in the ecology², evolution³ and physiology⁴ of plants. Despite its importance as a nutrient, very little is known about how AMF take up nitrogen and transfer it to their host plants⁵. Here we report the results of stable isotope labelling experiments showing that inorganic nitrogen taken up by the fungus outside the roots is incorporated into amino acids, translocated from the extraradical to the intraradical mycelium as arginine, but transferred to the plant without carbon. Consistent with this mechanism, the genes of primary nitrogen assimilation are preferentially expressed in the extraradical tissues, whereas genes associated with arginine breakdown are more highly expressed in the intraradical mycelium. Strong changes in the expression of these genes in response to nitrogen availability and form also support the operation of this novel metabolic pathway in the arbuscular mycorrhizal symbiosis.

The uptake of mineral nutrients from the soil by plants is greatly aided by mutualistic associations with mycorrhizal fungi, which grow into and extend out of the plant roots. Of these symbioses the arbuscular mycorrhizal one is the oldest, most anatomically intimate and ecologically widespread⁶. As well as benefiting plants by aiding phosphorus uptake from the soil⁷, AMF can take up and transfer significant amounts of nitrogen to their host plants⁵. The availability of nitrogen frequently limits plant growth, and depending on soil conditions nitrogen transfer by mycorrhizal fungi can represent a significant route of uptake by the plant^{5,8}. AMF have been strongly implicated in the transfer of nitrogen from one plant to another⁵, can increase the utilization of different forms of nitrogen by plants⁹ and have been shown to take up nitrogen directly and transfer it to host roots^{5,10,11}. However, despite the identification in AMF of enzymes and genes of primary nitrogen assimilation and catabolism (nitrate reductase, glutamine synthetase and glutamate dehydrogenase^{12–14}), we know very little about how nitrogen is transferred from fungus to plant. In particular, we do not know the form in which nitrogen is translocated within the fungus from the hyphae in the soil (extraradical mycelium) to the fungal structures within roots (intraradical mycelium), or the form in which nitrogen is transferred across the mycorrhizal interface to the plant. This ignorance limits our understanding both of underground nitrogen movement globally and of nutrient exchange in what is arguably the world's most important symbiosis.

In order to follow the uptake, assimilation and transfer of nitrogen in the arbuscular mycorrhizal symbiosis, we supplied isotopically labelled substrates to *in vitro* arbuscular mycorrhizal cultures of carrot (*Daucus carota* L.) roots colonized by *Glomus intraradices*.

When grown in divided Petri plates¹⁵, this model mycorrhiza excludes other microorganisms and prevents diffusion of non-volatile solutes between the compartments. Thus, nutrient transfer between the compartment in which the colonized roots grow and the compartment in which the extraradical mycelium (ERM) proliferates takes place only via uptake, metabolism and transport by the fungus^{16–20}. This model system develops normally with respect to fungal morphology and life cycle¹⁵, and functions in phosphorus uptake and transfer^{20,21}; it has also yielded considerable insight into carbon handling^{16–19,22–23}, and has been shown to take up and metabolize nitrogen^{10,14}.

The free amino acids of the ERM became highly labelled after this tissue was exposed to either ¹⁵NO₃[−] (Fig. 1a) or ¹⁵NH₄⁺ (not shown) added to the fungal compartment. This observation is consistent

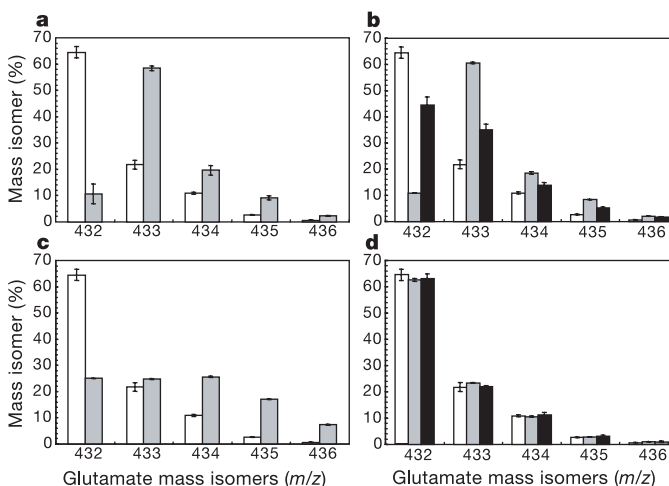


Figure 1 | Labelling of amino acids after supplying labelled nitrate or acetate to the extraradical mycelium. **a–d**, Mass isomer distributions of glutamate from free amino acids of the ERM (**a**, **c**), and free and protein amino acids of mycorrhizal roots (**b**, **d**) after supplying the ERM with ¹⁵N-labelled NO₃[−] (**a**, **b**) or ¹³C-labelled acetate (**c**, **d**) for 6 weeks. White bars represent the mass isomer distribution from an unlabelled glutamate standard, where most of the molecules (after derivatization) have a mass of 432 atomic mass units (a.m.u.), and the higher mass molecules have heavier isotope atoms at natural abundance levels. Grey bars represent glutamate from extracted free amino acids, and black bars represent glutamate from hydrolysed protein. In **b** and **c**, but not **d**, free glutamine and glutamine from protein have more molecules of higher mass than the unlabelled glutamine sample due to the metabolic incorporation of ¹³C and ¹⁵N atoms. Results are shown for glutamine as an abundant amino acid, but labelling patterns in other amino acids support the same conclusions. Error bars represent s.e.m.; *n* = 3.

¹Department of Chemistry and Biochemistry, New Mexico State University, Las Cruces, New Mexico 88003, USA. ²USDA-ARS Eastern Regional Research Center, 600 E. Mermaid Lane, Wyndmoor, Pennsylvania 19038, USA. ³Plant Biology Department, Michigan State University, East Lansing, Michigan 48824-1312, USA.

with the known uptake and assimilation of nitrogen into amino acids by the ERM^{10–11,14}. ¹⁵N labelling in the free amino acid pool from colonized roots was also very high (Fig. 1b), demonstrating nitrogen translocation to the colonized roots. Furthermore, the amino acids obtained by hydrolysis of total protein extracted from colonized roots were also significantly labelled (Fig. 1b). This finding is consistent with the transfer of nitrogen from the ERM to the host⁵. The protein extracted from mycorrhizal roots is expected to consist mostly of host protein, because the fungus and its protein represent a tiny proportion of the totals for these colonized roots (D.D.D. and H.B., unpublished results) as in all other arbuscular mycorrhizae examined⁶. The average ¹⁵N level in the nitrogen of root protein amino acids was $29.0 \pm 1.0\%$, which shows that close to one-third of the total nitrogen in the roots was provided by the fungus from inorganic nitrogen taken up by the fungal ERM in the fungal compartment. This is a high value when one considers that most of the root biomass was made before the labelling period. ¹⁵N labelling of the free amino acids of roots after supplying ¹⁵NO₃[−] or ¹⁵NH₄⁺ to the fungal compartment was very high, even when the nitrogen levels supplied in the root compartment were threefold higher than in this study (Y. S.-H., P.E.P. and D. Rolin, unpublished results), which shows that the uptake of nitrogen by the ERM and translocation to the mycorrhizal roots occurs whether the host roots are nitrogen limited or not.

In order to track the synthesis and fate of the carbon skeletons of amino acids made in the ERM, ¹³C₂ acetate was provided to the ERM compartment (sugars are not taken up by the ERM²²). Free amino acids in the ERM became highly labelled (Fig. 1c), with 34–76% of the molecules of the three abundant amino acids arginine, glutamate and aspartate having one or more ¹³C atoms. This finding indicates that the carbon skeletons of these amino acids are synthesized in the ERM, consistent with previous observations that acetate can enter central metabolism in the ERM via acetyl CoA and the glyoxylate cycle¹⁷. Free arginine extracted from the colonized roots after the ERM was exposed to ¹³C acetate was labelled between 6% and 18% (data not shown). This indicates that one or more amino acids are being translocated from the ERM to the intraradical mycelium (IRM). The alternative explanation that ¹³C from acetate is transferred to the IRM as storage lipid¹⁸ or carbohydrate and then made into amino acids is unlikely, because fungal compounds such as storage lipids and trehalose within the mycorrhizal roots did not become labelled in such experiments²³.

If nitrogen is transferred to the host root in the form of one or more free amino acids, then one would expect to see ¹³C labelling in

amino acids of isolated root proteins after ¹³C acetate supply to the ERM. However, protein from the colonized roots did not become detectably labelled with ¹³C under these conditions (Fig. 1d). The absence of ¹³C labelling in plant protein under conditions when large amounts of nitrogen are being transferred from fungus to plant indicates that nitrogen is transferred in inorganic form. To test the possibility that carbon is transferred with the nitrogen but that they are metabolically separated before nitrogen is incorporated into protein (as hypothesized for the ectomycorrhizal symbiosis²⁴), we analysed plant storage lipid and sucrose after ¹³C acetate was provided to the ERM, and found that neither of these metabolic pools became labelled²³. This indicates that carbon is not incorporated into other host carbon pools. In addition, we found that when ¹⁴C-labelled amino acids (alanine, arginine, glutamate, glutamine, or ornithine) were provided to uncolonized roots, between 10% and 20% of the amino acids that were taken up in one week were recovered in the soluble protein extracted from the roots.

The absence of ¹³C labelling in mycorrhizal root proteins indicates that the labelled free amino acids found in mycorrhizal root tissues after provisioning the ERM with ¹³C acetate are in the IRM, and that they are broken down to release inorganic nitrogen inside the fungus before transfer to the host. Analysis of the levels of free amino acids by high-performance liquid chromatography revealed that arginine is by far the most abundant fungal amino acid, (between 50 and 200 mM depending on developmental stage), representing >90% of the total free amino acids in the ERM. Arginine levels are also substantially higher in colonized compared with uncolonized roots ($54.2 \pm 19.3\%$ versus $10.9 \pm 4.8\%$ of free amino acids). This is consistent with a previous observation¹¹, and, together with the finding that arginine in the ERM is rapidly turned over (I. Jakobsen, C. Trujillo, P. Ambus, N. Requena and H. Egsgaard, unpublished data), suggests that arginine may be transported from the ERM to the IRM.

To test whether arginine is translocated from the ERM to the IRM, ¹³C₆ arginine was supplied to the ERM. After 6 weeks, 34% of the free arginine in the ERM and 33% of the free arginine in the colonized roots showed ¹³C₆ labelling (Fig. 2). The mass spectra show that the free arginine molecules in the colonized roots were either completely unlabelled (natural abundance mass isomer distribution) or labelled in all six carbon positions, thus indicating that arginine is transported intact from ERM to IRM. The absence of detectable ¹³C arginine labelling in proteins from the colonized roots in these experiments, despite the appearance of labelled free amino acids from colonized root samples, contrasts with the finding that

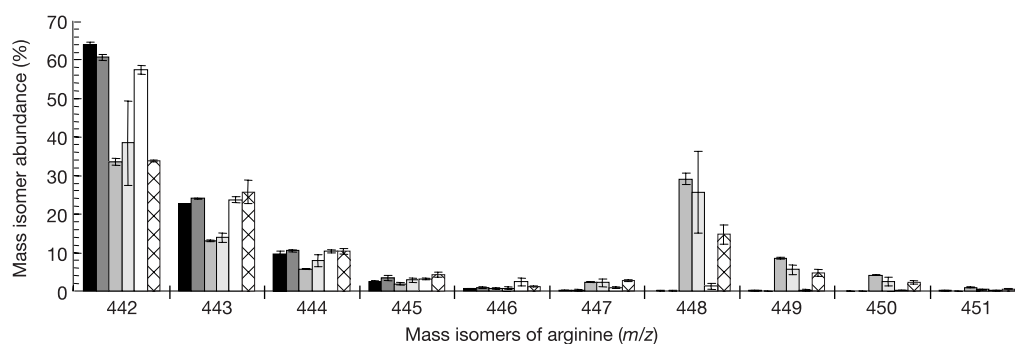


Figure 2 | Labelled arginine after addition of 2 mM ¹³C₆ arginine to the ERM compartment for 6 weeks. Mass isomer distributions were measured by mass spectrometry after extraction of free amino acids or hydrolysis of extracted soluble protein followed by derivatization (see Methods). Black bars, unlabelled arginine standard showing the natural abundance mass isomer distribution; dark grey bars, arginine extracted from unlabelled mycorrhizal root tissue; medium grey bars, arginine extracted from ERM

after labelling; light grey bars, free arginine from mycorrhizal roots after labelling; white bars, arginine from soluble protein of mycorrhizal roots after labelling; hatched bars, arginine from soluble protein of uncolonized roots after they were exposed to ¹³C₆ arginine (positive control, showing that if arginine is made available to the root tissue, it is detectable in root protein). Error bars represent s.e.m.; *n* = 3.

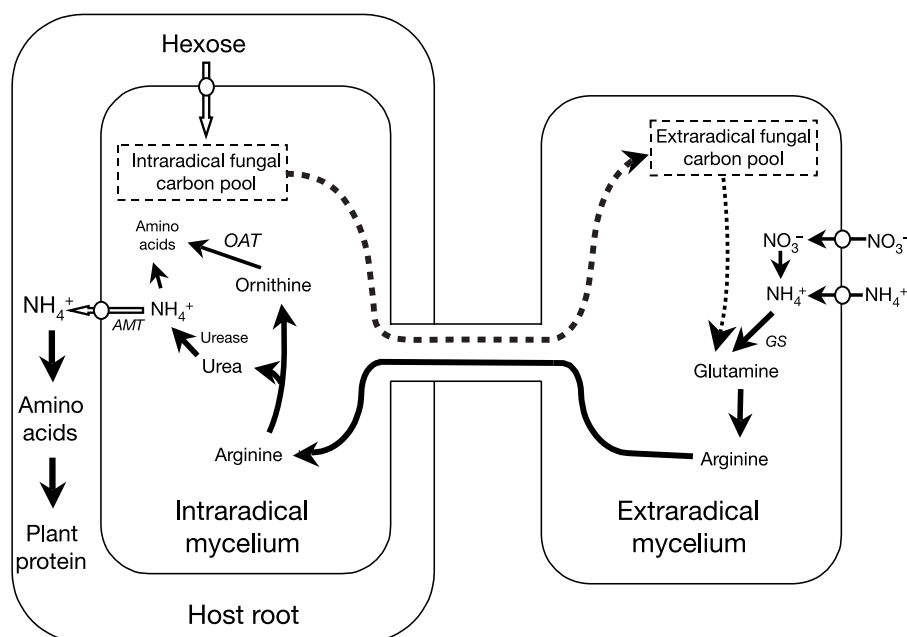


Figure 3 | Model of nitrogen movement in the arbuscular mycorrhizal symbiosis. Inorganic nitrogen is taken up by the fungal ERM and assimilated via nitrate reductase and the GS–GOGAT cycle. It is then converted into arginine, which is translocated along the coenocytic fungal hyphae from the ERM into the IRM. Arginine is broken down in the IRM,

releasing urea and ornithine, which are further broken down by the actions of urease and ornithine aminotransferase (OAT). Ammonia released from arginine breakdown passes to the host via ammonia channels (AMT). Amino acids from ornithine breakdown and/or NH_4^+ assimilation in the IRM may be catabolized within the IRM or translocated to the ERM.

roots supplied directly with labelled arginine incorporate significant levels into protein (see Fig. 2 and ^{14}C results described above). Supplying ^{15}N arginine to the ERM resulted in ^{15}N labelling of all the free amino acids in the root tissue, including those present at high levels in uncolonized roots and at low levels in the ERM (mass spectra not shown). Thus, the nitrogen but not the carbon of arginine is transferred from fungus to host across the host–fungus interface.

On the basis of these labelling patterns, we conclude that arginine synthesized in the ERM is translocated by the fungus from the ERM to the IRM, but that it is not transferred to the host; rather it is broken down in the IRM and nitrogen is transferred to the host as ammonium. This proposed mechanism, illustrated in Fig. 3, is consistent with both the suggestion that arginine may be involved in nitrogen transfer¹¹ and a hypothetical scheme we have previously presented²⁵. There is also evidence that arginine can bind to polyphosphate²⁶, which is the form of phosphorus thought to be translocated by the fungus⁶, suggesting a possible link between nitrogen and phosphorus movement.

The proposed mechanism (Fig. 3) predicts that enzyme activity required for nitrogen assimilation should be induced in the ERM and suppressed in the IRM when inorganic nitrogen is available to the ERM. We tested this prediction by measuring messenger RNA levels for key enzymes in ERM and IRM tissues by quantitative real-time polymerase chain reaction (PCR). Figure 4a shows that glutamine synthetase expression follows this prediction, particularly when NO_3^- is added to the ERM compartment, supporting previous findings that nitrogen assimilation occurs in the ERM via the glutamine synthetase–glutamate synthase (GS–GOGAT) pathway¹¹. Figure 4a also shows that expression of a putative NAD-dependent glutamate dehydrogenase (GDH) gene is downregulated in ERM tissue supplied with either NO_3^- or NH_4^+ in the ERM compartment, consistent with this enzyme having a catabolic role^{14,27,28}.

The proposed mechanism also requires that enzymes required for the breakdown of arginine are more active in the IRM than in the ERM in order to release ammonia for transfer to the host roots. Figure 4b shows that *G. intraradices* genes with high similarity to known ornithine aminotransferase, urease accessory protein and an

ammonium transporter are indeed preferentially expressed in the IRM, consistent with the predictions of the model. This was true regardless of whether NO_3^- or NH_4^+ was added to the ERM compartment, although the effect was greater for NH_4^+ treatments. The suggestion that NO_3^- is directly transferred from the ERM to the IRM and then to the plant²⁹ is not supported by the strong induction of glutamine synthetase transcripts in the ERM, observed when NO_3^- is provided to the ERM; similarly the high levels of labelling in ERM amino acids after $^{15}\text{NO}_3^-$ is supplied to the fungal compartment points to the assimilation of NO_3^- into amino acids in the ERM.

Our results show the operation of a metabolic route in which nitrogen is moved by the AMF from the soil to its host. This pathway

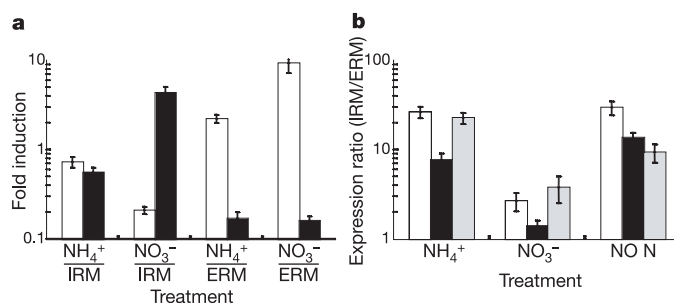


Figure 4 | Expression of primary nitrogen metabolic genes. Expression of putative nitrogen metabolic genes was measured in IRM and ERM by quantitative real-time PCR after 8 weeks of growth. Ribosomal protein S4 gene expression was used as the reference. **a**, The response of fungal gene expression to nitrogen availability. Levels of glutamine synthetase (white bars) and NAD-dependent glutamate dehydrogenase (black bars) mRNA are expressed as fold change relative to when no nitrogen was supplied to the fungal compartment. **b**, The ratio of intraradical to extraradical mycelial expression of urease accessory protein (white bars), ornithine aminotransferase (black bars) and ammonium transporter (grey bars). Relative expression values represent the ratios of normalized IRM to ERM mRNA levels. Error bars represent s.e.m.; $n = 3$.

consists of metabolic processes known to operate in fungi (the assimilation of inorganic nitrogen) together with a new variant of the urea cycle in which the anabolic and catabolic parts are separated by the long-distance translocation of arginine. The assimilation of nitrogen into arginine allows it to be moved in a concentrated form (four nitrogens per molecule) that is also non-toxic. Its potential to bind to polyphosphate might allow phosphorus translocation to also carry nitrogen. The use of the catabolic arm of the urea cycle allows the transfer of nitrogen to the host plant with minimal loss of carbon by the fungus. The existence of this pathway and the high flux of nitrogen through it indicate that the arbuscular mycorrhizal symbiosis can effectively transfer large amounts of nitrogen from the soil to plant roots. This means that the symbiosis may have a much more significant role in the global nitrogen cycle than has been widely believed⁶.

METHODS

Tissue culture. Root-inducing T-DNA-transformed carrot (*D. carota* clone DC2) roots colonized by *G. intraradices* (DAOM 197198) were grown²⁰ in bi-compartmented Petri plates²¹ as previously described²⁸. For labelling experiments, $\text{Ca}(\text{NO}_3)_2 \cdot 4\text{H}_2\text{O}$ was replaced with $180 \text{ mg l}^{-1} \text{ CaCl}_2 \cdot 2\text{H}_2\text{O}$ in the medium of both compartments. KNO_3 was increased to 100 mg l^{-1} in the root compartment, whereas in the fungal compartment the only nitrogen source was $^{15}\text{NO}_3^-$ (4 mM), $^{15}\text{NH}_4^+$ (4 mM), arginine-guanido- $^{15}\text{N}_2\text{HCl}$ (^{15}N , 98%) (2 mM), or $^{13}\text{C}_6$ (98%) arginine (2 mM). For ^{13}C labelling of the ERM, $^{13}\text{C}_2$ acetate was added to the fungal compartment. Filter-sterilized solutions of the labelled compounds (pH 6.0) were added to autoclaved medium gelled with 4 g l^{-1} phytalgel (Sigma). Tissue from two plates was combined per sample and three replicates were analysed per treatment. ERM and mycorrhizal roots were harvested after 6 weeks of labelling.

To test the possibility that nitrogen moved across the barrier, we conducted experiments in which ^{13}C -labelled glucose was supplied to the fungal or colonized root compartment. This substrate is highly mobile in the medium and is taken up and incorporated into sucrose by the roots but is not taken up by the extraradical fungal mycelium²². When supplied to the root compartment, labelling in host and fungal metabolites was measurable within 36 h by gas chromatography-mass spectrometry (GC-MS). When supplied to the fungal compartment, no labelling was detected in host or fungal metabolites at any time up to 12 weeks of incubation. Thus, there is no significant diffusion across the barrier connecting the two compartments. Experiments with ^{14}C acetate also showed that there was no significant movement of substrates between the compartments by diffusion.

For gene expression experiments, split plates with mycorrhizal roots growing on both sides were incubated at 24°C for 5 weeks, after which the roots and solidified minimal medium²⁸ from one compartment of each plate were transferred intact to a new split plate. Seventeen millilitres of fresh liquid M medium was then added to the empty half of each plate.

The liquid M medium contained 3.2 mM NH_4Cl (with no nitrate) or 3.2 mM NO_3^- (as KNO_3 and $\text{Ca}(\text{NO}_3)_2 \cdot 4\text{H}_2\text{O}$), or no nitrogen (KCl and $\text{CaSO}_4 \cdot 2\text{H}_2\text{O}$ replacing KNO_3 and $\text{Ca}(\text{NO}_3)_2 \cdot 4\text{H}_2\text{O}$, respectively). Only fungal hyphae grew over the barrier. ERM and mycorrhizal root samples were harvested 3 weeks after transfer.

Isolation of soluble and protein amino acids. Samples were stored at -80°C , ground with acid-washed sand in liquid nitrogen, and extracted twice with NH_4HCO_3 buffer (pH 8 with 0.2% NaN_3). After centrifugation the supernatant was lyophilized, re-suspended in NH_4HCO_3 buffer and dialysed twice against 40 ml of NH_4HCO_3 buffer at 4°C for 24 h using a dialysis membrane with a molecular mass cutoff of 2,000 Da (spectra/Por7 cellulose ester). Dialysed samples were pooled, lyophilized, stored at -20°C , then resuspended in 600 μl of 20 mM NH_4HCO_3 buffer.

Freshly dissolved proteases (2 μl aminopeptidase M, 2 μl of pronase E and 2 μl of carboxypeptidase Y) were added and the samples were incubated for 12 h at 30°C with constant shaking; fresh enzymes were added at 6 h. Samples were centrifuged for 10 min at 10,000g at 4°C , and the supernatants were lyophilized and re-suspended in 2 ml H_2O , then lyophilized again and re-suspended in 1 ml H_2O . The solution was loaded onto a cation exchange column (1 ml of DOWEX 50 *4-200, hydrogen form) and eluted with three 1-ml aliquots of 1 N NH_4OH . The eluate was lyophilized, re-suspended in 1 ml of water, acidified with 500 μl 1 N HCl, vortexed and then lyophilized. Free amino acids were purified as above either from the pooled dialysates (about 80 ml)—which were lyophilized, twice re-suspended in 2 ml water and lyophilized—or extracted MeOH as previously described²¹.

Amino acid analysis. Free amino acid abundances were determined using a Waters Pico-Tag amino acid analyser using the Pico-Tag method. The threshold for detection of amino acids in standard solutions was 30 pM of each amino acid per assay, corresponding to $<10 \text{ nmol g}^{-1}$ dry weight of tissue.

Derivatization of amino acids. Amino acid samples were dissolved in 20–50 μl of dry dimethyl formamide at room temperature. The dimethyl formamide was evaporated under N_2 gas, 75 μl of *N*-methyl-*N*-(*t*-butyldimethylsilyl)trifluoroacetamide containing 1% *N*-methyl-*N*-(*t*-butyldimethylchlorosilane) was added and the sample was heated for 30–50 min at 106°C .

GC-MS analysis of amino acid labelling. MS analyses were performed with a Trace 2000 gas chromatograph (Thermo Electron) equipped with a splitless injector, open-tubular column of 0.25- μm -thick BP-15 film (0.18 mm internal diameter, 30 m long, Agilent) interfaced to a Thermo Finnigan quadrupole mass detector (Thermo Electron). Temperatures were: injector, 290°C ; column, 50°C for 1 min after injection then increasing to 250°C at $10^\circ\text{C min}^{-1}$, then to 280°C at $25^\circ\text{C min}^{-1}$, and 280°C for 8 min; detector, 350°C . Carrier gas velocity was 1 ml min^{-1} . Peak identities were confirmed by GC-MS of authentic samples. The mass isomer distribution for each derivatized amino acid was measured from the intensity of ions having masses of 57 atomic mass units below the molecular ion, except for arginine whose *m*-188 ion was used. ^{15}N and ^{13}C labelling were measured from the intensities of signals with increased mass. Correction for background contributions to higher mass ions was made from the known N, C and Si isotopic natural abundances and confirmed with unlabelled standards and extracts.

The reliability of measuring fractional isotopic labelling was determined with samples of known labelling levels (natural abundance samples and glycine containing 0, 10%, 25%, 50%, 75% or 100% ^{15}N). Agreement between measured and predicted mass isomer distributions was always $>95\%$ and $>99\%$ for $>75\%$ of the repetitions.

Real-time RT-PCR. Total RNA was extracted from ERM tissues (Plant RNeasy kit, Qiagen) and from mycorrhizal roots (RNAqueous TM-Midi kit, Ambion) frozen in liquid nitrogen. RNAs were treated with DNase I (Ambion) and quantified by fluorescence (Ribogreen assay, Molecular Probes). Reverse transcription and PCR were performed with 1 ng and 2.5 ng of total RNA from ERM and IRM tissues, respectively (1-step master mix, Qiagen). Primers and probes (IDT, Synthege) were as follows: glutamine synthetase forward 5'-CCTCAAGGTCCCTATTATTGTTCTG-3', reverse 5'-ACGATAATGAGCTTCCA CAACGT-3', dual-labelled fluorogenic probes 5'-CGACCAAAAGCAACA TTCGCACCA-3'; GDH forward 5'-CCACTTATTGCATTTACGTCAAAGA-3', reverse 5'-CCCAGTCATCTCAGCAAGAGAA-3', dual-labelled fluorogenic probes 5'-CTTCTTCGCCATCCAATGGCAGC-3'; S4 ribosomal protein forward 5'-TCTTGTGAAGGTTGATGGCAA-3', reverse 5'-CGCCATTCTTTC GATCGA-3', dual-labelled fluorogenic probes 5'-TTCGAACCGATTCAACA TACCCTGCC-3'; UAP forward 5'-TTCCGCAGTCGTTGAATTGA-3', reverse 5'-CCTTCGACAATGCTTAAAAATTATCA-3', dual-labelled fluorogenic probes 5'-CACCAGACGCCTTCCAAGCACTCAAT-3'; OAT forward 5'-AGGGCTCAAGGTGCGTATGT-3', reverse 5'-ACTGCCGAATAGGCACAC AAA-3', dual-labelled fluorogenic probes 5'-TCCATATATTGTTGCC TTCTGGGTCCCA-3'. Gene-specific PCR products were monitored in an ABI PRISM 7700 Sequence Detection System (Applied Biosystems) using TaqMan probes. The *G. intraradices* S4 ribosomal protein was used for normalization. Expression values for IRM samples were compared with the expression in ERM tissues. The comparative ($\Delta\Delta C_T$) method was used to measure changes in gene expression³⁰. Errors are expressed as standard error of the mean (s.e.m.) of three independent biological triplicates determined as the mean of technical duplicates.

Received 6 October 2004; accepted 4 April 2005.

1. Read, D. J. & Perez-Moreno, J. Mycorrhizas and nutrient cycling in ecosystems — a journey towards relevance? *New Phytol.* **157**, 475–492 (2003).
2. Allen, M. F., Swenson, W., Querejeta, J. I., Egerton-Warburton, L. M. & Treseder, K. K. Ecology of Mycorrhizae: A conceptual framework for complex interactions among plants and fungi. *Annu. Rev. Phytopathol.* **41**, 271–303 (2003).
3. Redecker, D., Kodner, R. & Graham, L. E. Glomalean fungi from the Ordovician. *Science* **289**, 1920–1921 (2000).
4. Marschner, H. *Mineral Nutrition of Higher Plants* 2nd edn (Academic, London, 1995).
5. He, X.-H., Critchley, C. & Bledsoe, C. Nitrogen transfer within and between plants through common mycorrhizal networks (CMNs). *Crit. Rev. Plant Sci.* **22**, 531–567 (2003).
6. Smith, S. E. & Read, D. J. *Mycorrhizal Symbiosis* 2nd edn (Academic, San Diego, California, 1997).
7. Harrison, M. J. & Van Buuren, M. L. A phosphate transporter from the mycorrhizal fungus *Glomus versiforme*. *Nature* **378**, 626–629 (1995).

8. Ames, R. N., Reid, C. P. P., Porter, L. K. & Cambardella, C. Hyphal uptake and transport of nitrogen from two ^{15}N -labelled sources by *Glomus mosseae*, a vesicular-arbuscular mycorrhizal fungus. *New Phytol.* **95**, 381–396 (1983).
9. Hodge, A., Campbell, C. D. & Fitter, A. H. An arbuscular mycorrhizal fungus accelerates decomposition and acquires nitrogen directly from organic material. *Nature* **413**, 297–299 (2001).
10. Bago, B., Vierheilig, H., Piché, Y. & Azcon-Aguilar, C. Nitrate depletion and pH changes induced by the extraradical mycelium of the arbuscular mycorrhizal fungus *Glomus intraradices* grown in monoxenic culture. *New Phytol.* **133**, 273–280 (1996).
11. Johansen, A., Finlay, R. D. & Olsson, P. A. Nitrogen metabolism of external hyphae of the arbuscular mycorrhizal fungus *Glomus intraradices*. *New Phytol.* **133**, 705–712 (1996).
12. Kaldorf, M., Zimmer, W. & Bothe, H. Genetic-evidence for the occurrence of assimilatory nitrate reductase in arbuscular mycorrhizal and other fungi. *Mycorrhiza* **5**, 23–28 (1994).
13. Breuninger, M., Trujillo, C. G., Serrano, E., Fischer, R. & Requena, N. Different nitrogen sources modulate activity but not expression of glutamine synthetase in arbuscular mycorrhizal fungi. *Fungal Genet. Biol.* **41**, 542–552 (2004).
14. Toussaint, J. P., St-Arnaud, M. & Charest, C. Nitrogen transfer and assimilation between the arbuscular mycorrhizal fungus *Glomus intraradices* Schenck & Smith and Ri T-DNA roots of *Daucus carota* L. in an *in vitro* compartmented system. *Can. J. Microbiol.* **50**, 251–260 (2004).
15. St-Arnaud, M., Hamel, C., Vimard, B., Caron, M. & Fortin, J. A. Enhanced hyphal growth and spore production of the arbuscular mycorrhizal fungus *Glomus intraradices* in an *in vitro* system in the absence of host roots. *Mycol. Res.* **100**, 328–332 (1996).
16. Bago, B., Pfeffer, P. E. & Shachar-Hill, Y. Carbon metabolism and transport in arbuscular mycorrhizas. *Plant Physiol.* **124**, 949–957 (2000).
17. Lammers, P. J. et al. The glyoxylate cycle in an arbuscular mycorrhizal fungus: gene expression and carbon flow. *Plant Physiol.* **127**, 1287–1298 (2001).
18. Bago, B. et al. Translocation and utilization of fungal lipid in the arbuscular mycorrhizal symbiosis. *Plant Physiol.* **128**, 108–124 (2002).
19. Bago, B. et al. Carbon export from arbuscular mycorrhizal roots involves the translocation of carbohydrate as well as lipid. *Plant Physiol.* **131**, 1496–1507 (2003).
20. Bücking, H. & Shachar-Hill, Y. Phosphate uptake, transport and transfer by the AM fungus *Glomus intraradices* is stimulated by increased carbohydrate availability. *New Phytol.* **165**, 899–912 (2005).
21. Nielsen, J. S., Jøner, E. J., Declerck, S., Olsson, S. & Jakobsen, I. Phospho-imaging as a tool for visualization and noninvasive measurement of P transport dynamics in arbuscular mycorrhizas. *New Phytol.* **154**, 809–819 (2002).
22. Pfeffer, P. E., Douds, D. D., Bécard, G. & Shachar-Hill, Y. Carbon uptake and the metabolism and transport of lipids in an arbuscular mycorrhiza. *Plant Physiol.* **120**, 587–598 (1999).
23. Pfeffer, P. E., Douds, D. D., Bücking, H., Schwartz, D. P. & Shachar-Hill, Y. The fungus does not transfer carbon to or between roots in an arbuscular mycorrhizal symbiosis. *New Phytol.* **163**, 617–627 (2004).
24. Martin, F., Stewart, G. R., Genetet, I. & Le Tacon, F. Assimilation of $^{15}\text{NH}_4^+$ by beech (*Fagus sylvatica* L.) ectomycorrhizas. *New Phytol.* **102**, 85–94 (1986).
25. Bago, B., Pfeffer, P. E. & Shachar-Hill, Y. Could the urea cycle be translocating nitrogen in the arbuscular mycorrhizal symbiosis? *New Phytol.* **149**, 4–8 (2001).
26. Martin, F. ^{15}N -NMR studies of nitrogen assimilation and amino acid biosynthesis in the ectomycorrhizal fungus *Cenococcum graniforme*. *FEBS Lett.* **182**, 350–354 (1985).
27. Cliquet, J. B. & Stewart, G. R. Ammonia assimilation in *Zea-mays* L. infected with a vesicular-arbuscular mycorrhizal fungus *Glomus fasciculatum*. *Plant Physiol.* **101**, 865–871 (1993).
28. Vallorani, L. et al. Biochemical and molecular characterization of NADP glutamate dehydrogenase from the ectomycorrhizal fungus *Tuber borchii*. *New Phytol.* **154**, 779–790 (2002).
29. Hildebrandt, U., Schmelzer, E. & Bothe, H. Expression of nitrate transporter genes in tomato colonized by an arbuscular mycorrhizal fungus. *Physiol. Plant.* **115**, 125–136 (2002).
30. Winer, J., Jung, C. K. S., Shackel, I. & Williams, P. M. Development and validation of real-time quantitative reverse-transcriptase-polymerase chain reaction for monitoring gene expression in cardiac myocytes *in vitro*. *Anal. Biochem.* **270**, 41–49 (1999).

Acknowledgements We thank A. Abdul-Wakeel and D. Schwartz for their technical assistance, and acknowledge the financial support of this project from USDA-NRICGP.

Author Contributions Experimental work was carried out by J.A., J.W.A., H.B., D.D.D., M.G., H.J. and PEP; experimental design was by H.B., M.G., P.J.L., P.E.P. and Y.S.-H.; data analysis was carried out by M.G., H.J., P.J.L., P.E.P. and Y.S.-H.; and project planning was by Y.S.-H.

Author Information Accession numbers (GenBank) for gene sequences used to design the TaqMan assays are as follows: GS (DQ063587), GDH (AY745984), UAP (CV186300), OAT (BI452207) and S4 RP (BI452093). Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to P.J.L. (plammers@nmsu.edu) or P.E.P. (ppfeffer@arserrc.gov).

LETTERS

Plant sesquiterpenes induce hyphal branching in arbuscular mycorrhizal fungi

Kohki Akiyama^{1,2}, Ken-ichi Matsuzaki¹ & Hideo Hayashi¹

Arbuscular mycorrhizal (AM) fungi form mutualistic, symbiotic associations with the roots of more than 80% of land plants¹. The fungi are incapable of completing their life cycle in the absence of a host root. Their spores can germinate and grow in the absence of a host, but their hyphal growth is very limited. Little is known about the molecular mechanisms that govern signalling and recognition between AM fungi and their host plants. In one of the first stages of host recognition, the hyphae of AM fungi show extensive branching in the vicinity of host roots before formation of the appressorium^{2–4}, the structure used to penetrate the plant root. Host roots are known to release signalling molecules that trigger hyphal branching^{5–7}, but these branching factors have not been isolated. Here we have isolated a branching factor from the root exudates of *Lotus japonicus* and used spectroscopic analysis and chemical synthesis to identify it as a strigolactone, 5-deoxy-strigol. Strigolactones are a group of sesquiterpene lactones, previously isolated as seed-germination stimulants for the parasitic weeds *Striga* and *Orobanchae*⁸. The natural strigolactones 5-deoxy-strigol, sorgolactone and strigol, and a synthetic analogue, GR24, induced extensive hyphal branching in germinating spores of the AM fungus *Gigaspora margarita* at very low concentrations.

AM fungi are ancient, obligate symbionts in the phylum Glomeromycota⁹. The critical developmental step in their life cycle is hyphal branching, which helps them to ensure contact with the host root and the establishment of symbiosis. Branching factor (BF) is hypothesized to be a plant signal molecule needed to trigger the hyphal morphogenesis that precedes successful root colonization^{5,6}. Dialysis membranes have been used to determine that the BF exuded from growing roots of *Ocimum basilicum*, which elicits hyphal branching of the AM fungus *Glomus mosseae*, is a low-molecular-weight compound (less than 500 Da)⁵. The development of an *in vitro* bioassay for hyphal branching in germinating spores of the genus *Gigaspora*¹⁰ has facilitated the analysis of the chemical characteristics and distribution of BF in the plant kingdom^{6,7}. BF was present in root exudates of all the mycotrophic plants tested, but absent in those of non-host plants. The discovery that the active compound is partitioned into ethyl acetate from an aqueous root exudate⁶, and is retained on C₁₈ reverse-phase resin⁷, indicates that BF is a lipophilic compound. The presence of several BFs in host plants was suggested by C₁₈ reverse-phase column and preparative thin-layer chromatography⁷. Flavonoids have been ruled out as BF candidates because root exudates of maize mutants deficient in chalcone synthase show comparable activity to those of the wild type⁶. Root exudates from plants grown under phosphate (Pi)-limited conditions are more active than those from plants with sufficient Pi nutrition, suggesting that the production of BF in roots and its exudation are regulated by Pi availability⁷. The identity of BF has been the subject of a considerable amount of research, but

purification of the root metabolites has been hampered by the extremely low concentrations produced and exuded by host roots¹¹ and their relative instability (K.A., unpublished data). Here we report the isolation and structural identification of BF from root exudates of the model legume *L. japonicus*.

Plants were grown hydroponically under low-Pi conditions (35 μ M). Lipophilic compounds released into the hydroponic solution were extracted with ethyl acetate and then tested for hyphal branching activity in germinating spores of *G. margarita*. Using the paper disk diffusion method, the ethyl acetate extract elicited strong hyphal branching at 15 μ g per disk (Fig. 1a, b). The extract showed activity at concentrations as low as 1.9 μ g per disk. The ethyl acetate extract was further divided into ethyl acetate-soluble acidic, neutral and basic fractions. The activity was found only in the neutral fraction, indicating that BF from *L. japonicus* is a neutral compound (Fig. 1c, d).

To isolate the neutral active compound, we used activated charcoal

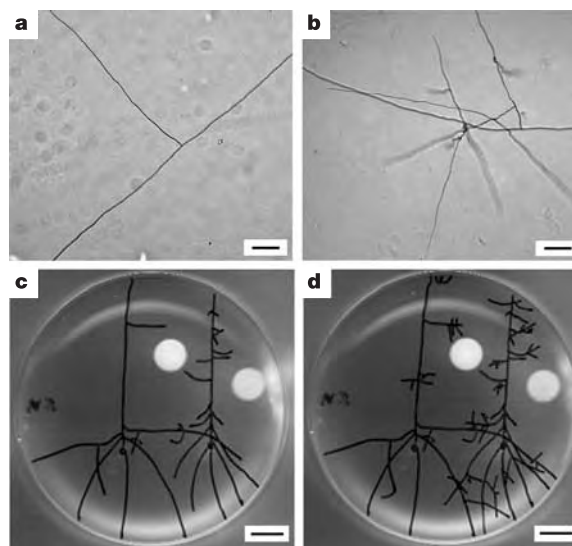


Figure 1 | Hyphal branching of *G. margarita* induced by lipophilic fractions from root exudates of *L. japonicus* using the paper disk diffusion method. **a**, Control hypha (70% ethanol in water). **b**, Hyphal branching from a secondary hypha upon treatment with the ethyl acetate extracts (15 μ g per disk). Scale bars, 300 μ m. **c**, **d**, Tracings of hyphal branches directly traced on a Petri dish at 0 h (**c**) and 24 h (**d**) after treatment with the ethyl acetate soluble-neutral fraction (15 μ g per disk). Extensive formation of hyphal branches from secondary and primary hyphae was induced. Scale bars, 6 mm.

¹Division of Applied Biological Chemistry, Graduate School of Agriculture and Biological Sciences, Osaka Prefecture University, Sakai, Osaka 599-8531, Japan. ²Core Research for Evolutional Science and Technology (CREST), Japan Science and Technology Agency, 4-1-8 Honcho, Kawaguchi, Saitama 332-0112, Japan.

to extract BF from the hydroponic solution. The solution was continuously circulated through an activated charcoal cartridge. Active compounds adsorbed on the charcoal were eluted with acetone. The ethyl acetate-soluble, neutral fraction prepared from the acetone eluate was first chromatographed on a silica gel column by eluting stepwise from *n*-hexane to ethyl acetate. Activity was found in 40% ethyl acetate eluate. The active fraction was further purified by high-performance liquid chromatography (HPLC) on a semi-preparative C_{18} reverse-phase column, resulting in the isolation of a BF eluted as a single peak at 22.3 min (Fig. 2). None of the other peaks was active in the assay.

We subjected the BF, obtained as a colourless, amorphous solid, to spectroscopic analysis. The ultraviolet spectrum showed an absorption maximum at 234 nm, suggesting an α , β -unsaturated carbonyl chromophore. The infrared spectrum exhibited strong absorption bands indicative of γ -lactones ($1,784$, $1,744\text{ cm}^{-1}$) and enol ether ($1,683\text{ cm}^{-1}$). The electron impact (EI) mass spectrum showed a weak molecular ion peak at m/z 330, with the base peak at m/z 97. The spectrum indicated cleavages from m/z 330 to peaks at m/z 315 [$M^+ - \text{CH}_3$], 233 [$M^+ - 97$], 215 [$M^+ - 97 - \text{H}_2\text{O}$], and 187 [$M^+ - 97 - \text{H}_2\text{O} - \text{CO}$] (Fig. 3a). The proton-nuclear magnetic resonance (^1H -NMR) spectrum displayed signals for two singlet methyls (δ_{H} 1.08, 1.10), an allylic methyl (δ_{H} 2.01, triplet, $J = 1.5\text{ Hz}$), an oxymethine (δ_{H} 5.49, broad doublet, $J = 8.0\text{ Hz}$), and an allylically coupled olefinic proton (δ_{H} 7.39, doublet, $J = 2.7\text{ Hz}$). Two methine protons were observed at δ_{H} 6.12 and 6.90 as a multiplet, respectively. Taken together, these data strongly suggested that the BF is structurally closely related to strigolactones, a group of sesquiterpene lactones previously identified as seed germination stimulants for the parasitic weeds *Striga* and *Orobancha*. To date, five natural strigolactones have been isolated (Fig. 3e, f)^{12–17}, and a number of structural analogues have been synthesized (Fig. 3g)¹⁸. The spectral data of the BF were in good agreement with those reported for a synthetic derivative of strigol, ($\pm$)-5-deoxy-strigol¹⁹, which has not been isolated from any natural source²⁰. Strigol itself is the first natural strigolactone to be isolated from the false host cotton^{12,13}. To confirm the chemical structure of the BF, ($\pm$)-5-deoxy-strigol was synthesized according to the previously reported procedures^{19,21,22}. Retention times on HPLC and all the spectral data for the synthetic compound were identical to those of the natural compound (Fig. 3b and Supplementary Table 1). Thus, the BF isolated from *L. japonicus* was identified as 5-deoxy-strigol. The circular dichroism (CD) spectrum of the natural compound was superimposable on that reported for the natural strigolactones (+)-strigol and (+)-sorgolactone (Fig. 3c), confirming that the absolute configuration of the natural 5-deoxy-strigol is the same as that of the naturally occurring strigolactones (that is, 3a(R), 8b(S) and 2'(R)) (Fig. 3d)^{19,21,23}. Little is known about the biosynthetic pathway for strigolactones in plants owing to the extremely low concentrations of highly active compounds that are produced by and

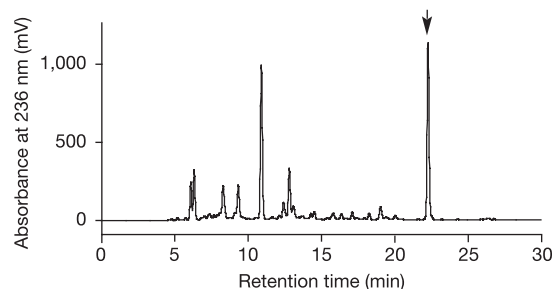


Figure 2 | Purification of BF by HPLC on a semi-preparative C_{18} reverse-phase column. A 40% ethyl acetate eluate was obtained from the first silica gel chromatography of the ethyl acetate-soluble neutral fraction from root exudates of *L. japonicus*. A peak eluted at 22.3 min (arrow).

exuded from the roots. Isolated as a natural product for the first time, 5-deoxy-strigol can be converted to strigol and orobanchol¹⁷ by hydroxylation at C-5 and C-4, respectively, suggesting that this compound is a branching point in the strigolactone biosynthesis.

To discover whether the natural 5-deoxy-strigol elicited hyphal branching, we tested it on *G. margarita* at concentrations ranging from 1 ng to 1 pg per disk. Natural BF elicited the same hyphal branching as that observed for the ethyl acetate-soluble neutral fractions prepared from root exudates of *L. japonicus* at 1 ng–30 pg

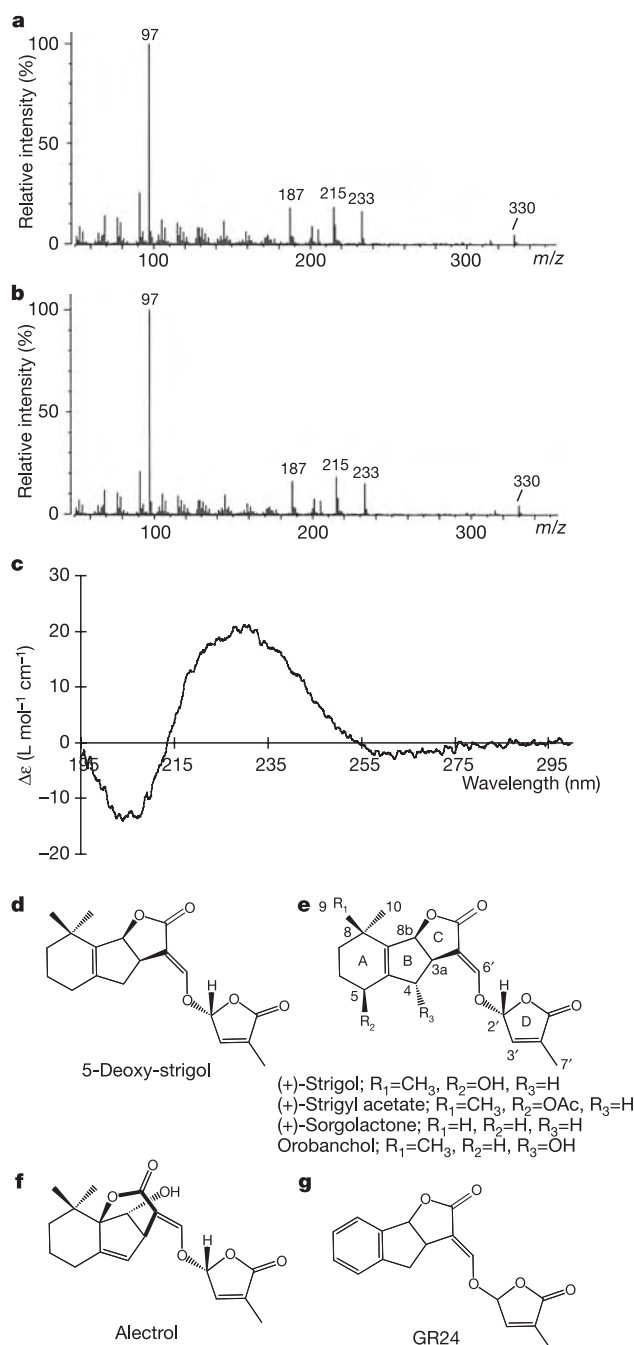


Figure 3 | Spectroscopic analysis of BF isolated from *L. japonicus*. **a**, **b**, EI mass spectra of BF from *L. japonicus*, identified as 5-deoxy-strigol (**a**), and synthetic ($\pm$)-5-deoxy-strigol (**b**). **c**, CD spectrum of BF in acetonitrile. **d–f**, Chemical structures of BF isolated from *L. japonicus*, natural strigolactones and a synthetic strigolactone analogue. **d**, BF, identified as 5-deoxy-strigol; **e**, four natural strigolactones; **f**, aleictrol (the structure has not been confirmed); **g**, GR24.

per disk (Fig. 4a). This compound was still active at 10–3 pg per disk, inducing extensive secondary hyphae from the primary hypha.

To study structure–activity relationships of the BF, ($\pm$)-5-deoxy-strigol, ($\pm$)-sorgolactone, (+)-strigol and the synthetic strigolactone analogue GR24 were tested for activity at concentrations ranging from 100 ng to 30 pg per disk. At all the concentrations tested (100 ng–30 pg per disk), ($\pm$)-5-deoxy-strigol exhibited activity (Fig. 4b). Sorgolactone, a mono-8-demethyl derivative of 5-deoxy-strigol, is the third natural strigolactone to be isolated from sorghum root exudates¹⁴. Racemic sorgolactone prepared by organic synthesis²¹ was also highly active on the AM fungus (Fig. 4c). The activity of this compound was comparable to that of ($\pm$)-5-deoxy-strigol (100 ng–30 pg per disk). Isolated from an aseptically root culture of a Chinese medicinal plant *Menispermum dauricum*²⁴, (+)-strigol also induced hyphal branching at 1–100 ng per disk (Fig. 4d), though its activity at lower than 1 ng per disk has not yet been evaluated. GR24 was less active than the natural strigolactones (100–1 ng per disk), even considering that the GR24 used in this study was an equimolar mixture of two racemic diastereomers. Strigolactones are highly active on parasitic weeds, inducing 50% seed germination at picomolar concentrations^{14,15}. The CD part (see Fig. 3e) of strigolactone molecules was shown to be responsible for stimulating germination²⁵. A proposed molecular mechanism for germination stimulation involves the addition of a nucleophilic species, present at a putative receptor site, to the enol ether carbon double bond in a Michael fashion, followed by elimination of the D ring. The inherent instability of strigolactones in the presence of nucleophilic agents can also be understood by this mechanism. Taken with our observations that all the strigolactones tested had a potent effect on the fungus, and that their activity drastically decreased after concentrating a solution of BF dissolved in nucleophilic solvents such as methanol and its aqueous mixture, it appears that the CD part is also essential for the effect of strigolactones on the AM fungus.

Strigolactones have been isolated from the root exudates of a variety of plants⁸, including the monocots sorghum, maize and proso millet, and the dicots cotton, cowpea, red clover and *M. dauricum*. The isolation of (+)-strigol from an aseptically root culture of *M. dauricum* has unambiguously demonstrated that this strigolactone is of plant origin²⁴. Although it has been suggested that

strigolactones are more widely distributed in the plant kingdom, the isolation and characterization of strigolactones in root exudates, as in the case of this BF, have been hampered by the extremely low concentrations produced and exuded by host roots and their relative instability. It is noteworthy that *Arabidopsis thaliana*, which belongs to the nonmycotrophic family Brassicaceae, has been found to produce seed germination stimulants, but at much lower concentrations than in the crop plants carrot and tobacco²⁶, which have symbiotic fungi. The broad distribution of strigolactones in the plant kingdom and their levels in plant root exudates are consistent with the host specificity of AM fungi. Analysis using HPLC connected to tandem mass spectrometry²⁷ indicates the presence of several novel strigolactones, in addition to the known ones, in the root exudates of soybean, carrot, tomato and pea (K. Yoneyama, personal communications). We also isolated two novel BF-active compounds from the root exudates of sorghum. Intense, characteristic fragment peaks at m/z 97 in the EI mass spectra showed these to be strigolactone derivatives (K.A., K.M. and H.H., unpublished data).

Across large swathes of the world, the parasitic weeds *Striga* and *Orobancha* are among the most damaging agricultural pests⁸. Most species are obligate parasites incapable of completing their life cycle in the absence of a host. It has been suggested that strigolactones play a role in the natural ecological host–parasite relation because (+)-strigol was first isolated from the false host cotton. Given that production of strigolactones by red clover roots is stimulated under low-Pi conditions, and that parasitic weeds prevail in areas with limited Pi availability in the soil²⁸, it is tempting to speculate that parasitic weeds might find potential hosts by detecting strigolactones (which are released from plant roots under conditions of Pi deficiency in communication with AM fungi). The purification and characterization of plant symbiotic signals open up new ways to study the molecular basis of plant–AM fungus interactions. Availability of BF and its analogues will facilitate the detailed analysis of the molecular, physiological and morphological responses of AM fungi to this signal molecule. The use of the model legume *L. japonicus*²⁹ will further our understanding of the BF biosynthetic pathway and its regulation. Molecular and chemical analysis of BF-mediated events during the development of the AM symbiosis between *L. japonicus* and *G. margarita* will provide insights into the chemical communication between plants and AM fungi. This could also provide a new strategy for the management and control of beneficial fungal symbionts and devastating parasitic weeds in both agriculture and natural ecosystems.

METHODS

Spectroscopic methods. Infrared spectra were recorded with a JASCO FT/IR-460plus spectrometer, and ultraviolet spectra were measured with a Hitachi U-3210 instrument. Mass spectra were recorded on a JEOL JMS-700 instrument. ¹H-NMR spectra were obtained with a JEOL JNM-AL400 NMR spectrometer. Chemical shifts were referenced to the solvent peak (CDCl₃ δ_H 7.24) as an internal standard. CD spectra were measured with a JASCO J-820 spectropolarimeter.

Hyphal branching assay. Effects on the hyphal branching of AM fungi were evaluated *in vitro* by the paper disk diffusion method. Spores of *G. margarita* Becker & Hall (CGC1411, Central Glass Co.), surface-sterilized with 0.2% NaClO and 0.05% Triton X-100 (and rinsed with sterilized distilled water), were inserted into a 0.2% Phytigel gel (Sigma-Aldrich) containing 3 mM MgSO₄ in 60-mm plastic Petri dishes. The dishes were incubated vertically for 5–7 days in a 2% CO₂ incubator at 32 °C. Secondary hyphae emerging from a primary hypha, which grew upward in a negative geotropic manner in the gel, were used for assay. Test samples were first dissolved in acetone then diluted with 70% ethanol in water. The concentration of test sample solutions of natural 5-deoxy-strigol was adjusted with reference to the calibration of synthetic ($\pm$)-5-deoxy-strigol in HPLC analysis. Paper disks (6 mm in diameter; ADVANTEC) loaded with 15 μ l of test sample solution were placed in front of the tips of the secondary hyphae. The control was 70% ethanol in water. The hyphal branch patterns were observed 24 h after treatment. The sample was scored as positive for hyphal branching if new hyphal branches formed from the treated secondary hyphae or primary hyphae located proximal to the paper disks.

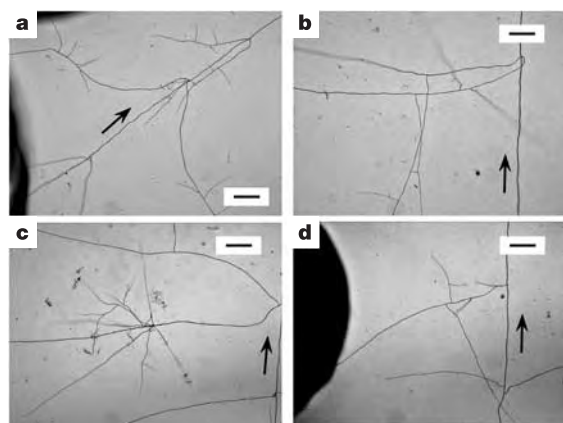


Figure 4 | Hyphal branching activity of BF and natural strigolactones on *G. margarita*. **a**, Natural BF, 5-deoxy-strigol isolated from *L. japonicus* (30 pg per disk). **b**, Synthetic BF, ($\pm$)-5-deoxy-strigol (3 ng per disk). **c**, ($\pm$)-Sorgolactone prepared by organic synthesis (10 ng per disk). **d**, (+)-Strigol isolated from *M. dauricum* root culture (3 ng per disk). Extensive emergence of tertiary, fourth and fifth hyphae from secondary hyphae was induced by treatment with BF or strigolactones. Arrows indicate direction of growth of primary hyphae. Control hyphae, which were treated with 70% ethanol–water, formed no hyphal branches, as shown in Fig. 1a. Scale bars, 600 μ m.

The assay was repeated at least twice, using between three and five dishes for each concentration.

Hydroponic culture. Seeds (~2,000) of *L. japonicus* B-129 Gifu were sown on vermiculite in a plastic tray with a meshed bottom and grown for 11 days in a growth cabinet with a 16/8-h day/night cycle at 25/22 °C. The tray was placed in a plastic vessel filled with ~20 l M medium (pH 6.8) without sucrose and vitamins³⁰. The solution was continuously aerated with an aquarium water pump. At 5 weeks after sowing, an activated charcoal cartridge was placed in the water intake port of the pump to extract BF from the solution. The hydroponic solution was used for the solvent fractionation experiments.

Solvent fractionation of BF. The hydroponic solution was adjusted to pH 2 using HCl and extracted with ethyl acetate. The ethyl acetate phase was partitioned with 0.2 M K₂HPO₄ (pH 9). The remaining ethyl acetate phase was dried over Na₂SO₄ and concentrated to give a neutral fraction. The first aqueous phase was adjusted to pH 12 using NaOH and extracted with ethyl acetate to give an ethyl acetate-soluble basic fraction. The second 0.2 M K₂HPO₄ phase was acidified to pH 2.0 and extracted with ethyl acetate to give an ethyl acetate-soluble acidic fraction.

Purification of BF (5-deoxy-strigol). After three days in the water intake port of the pump, the activated charcoal cartridge was eluted with acetone. The ethyl acetate-soluble neutral fraction (4 mg) prepared from the acetone eluate was chromatographed on a silica gel column (Wakogel C-200, Wako Pure Chemical Industries) and eluted stepwise with solvents of increasing polarity from *n*-hexane to ethyl acetate, with 10% increments. The fraction eluted with 40% ethyl acetate was chromatographed on an Inertsil ODS-3 HPLC column (φ10 × 250 mm, 5 μm; GL Sciences) employing a linear gradient elution within 30 min from 40% acetonitrile in water to 100% acetonitrile at a flow rate of 3.8 ml min⁻¹. Compounds eluted from the column were monitored at 236 nm. At 22.3 min, 5-deoxy-strigol eluted as a single peak. This fraction was evaporated to water *in vacuo* and extracted three times with ethyl acetate. After removal of the organic solvent *in vacuo*, 5-deoxy-strigol (18 μg) was obtained as a colourless, amorphous solid.

5-Deoxy-strigol: ultraviolet (acetonitrile) λ_{max} 234 nm; infrared (KBr) ν_{max} 1,784, 1,744, 1,683, 1,340, 1,024 cm⁻¹; ¹H-NMR (400 MHz, CDCl₃) δ 1.08 (3H, s, H-9 or H-10), 1.10 (3H, s, H-9 or H-10), 2.01 (3H, t, *J* = 1.5 Hz, H-7'), 5.49 (1H, br.d, *J* = 8.0 Hz, H-8b), 6.12 (1H, m, H-2'), 6.90 (1H, m, H-3'), 7.39 (1H, d, *J* = 2.7 Hz, H-6'); EIMS 70 eV, *m/z* (rel. int) 330 [M]⁺(4), 315 (2), 233 (16), 216 (9), 215 (18), 205 (7), 201 (9), 187 (18), 97 (100); CD (acetonitrile) λ_{max} (Δε) 262 (-2.86), 230 (21.2) nm.

Synthesis of (±)-5-deoxy-strigol. The racemic ABC ring was synthesized²¹, starting from 2,2-dimethyl-cyclohexanone²². Ester condensation of the racemic ABC ring with ethyl formate followed by alkylation with racemic 4-bromo-2-methyl-2-buten-4-olide provided (±)-5-deoxy-strigol and its 2'-epimer¹⁹. (±)-5-Deoxy-strigol was purified by a silica gel column (Wakogel C-200, *n*-hexane-ethyl acetate stepwise) and semi-preparative HPLC (Inertsil ODS-3, 70% acetonitrile in water).

Received 1 March; accepted 4 April 2005.

- Smith, S. E. & Read, D. J. *Mycorrhizal Symbiosis* (Academic, San Diego, 1997).
- Mosse, B. & Hepper, C. Vesicular-arbuscular mycorrhizal infections in root organ cultures. *Physiol. Plant Pathol.* **5**, 215–223 (1975).
- Giovannetti, M., Sbrana, C., Avio, L., Citernesi, A. S. & Logi, C. Differential hyphal morphogenesis in arbuscular mycorrhizal fungi during preinfection stages. *New Phytol.* **125**, 587–593 (1993).
- Giovannetti, M., Sbrana, C. & Logi, C. Early process involved in root recognition by arbuscular mycorrhizal fungi. *New Phytol.* **127**, 703–709 (1994).
- Giovannetti, M., Sbrana, C., Silvia, A. & Avio, L. Analysis of factors involved in fungal recognition response to host-derived signals by arbuscular mycorrhizal fungi. *New Phytol.* **133**, 65–71 (1996).
- Buee, M., Rossignol, M., Jauneau, A., Ranjeva, R. & Bécard, G. The pre-symbiotic growth of arbuscular mycorrhizal fungi is induced by a branching factor partially purified from plant root exudates. *Mol. Plant Microbe Interact.* **13**, 693–698 (2000).
- Nagahashi, G. & Douds, D. D. Partial separation of root exudate compounds and their effects upon the growth of germinated spores of AM fungi. *Mycol. Res.* **104**, 1453–1464 (2000).
- Bouwmeester, H. J., Matusova, R., Zhongkui, S. & Beale, M. H. Secondary metabolite signalling in host-parasitic plant interactions. *Curr. Opin. Plant Biol.* **6**, 358–364 (2003).
- Schussler, A., Schwarzott, D. & Walker, C. A new fungal phylum, the Glomeromycota: phylogeny and evolution. *Mycol. Res.* **105**, 1413–1421 (2001).
- Nagahashi, G. & Douds, D. D. A rapid and sensitive bioassay with practical application for studies on interactions between root exudates and arbuscular mycorrhizal fungi. *Biotechnol. Tech.* **13**, 893–897 (1999).
- Tamasloukht, M. *et al.* Root factors induce mitochondrial-related gene expression and fungal respiration during the developmental switch from asymbiosis to presymbiosis in the arbuscular mycorrhizal fungus *Gigaspora rosea*. *Plant Physiol.* **131**, 1468–1478 (2003).
- Cook, C. E., Whichard, L. P., Turner, B., Wall, M. E. & Egley, G. H. Germination of witchweed (*Striga lutea* Lour.): Isolation and properties of a potent stimulant. *Science* **154**, 1189–1190 (1966).
- Cook, C. E. *et al.* Germination stimulants. II. The structure of strigol—A potent seed germination stimulant for witchweed (*Striga lutea* Lour.). *J. Am. Chem. Soc.* **94**, 6198–6199 (1972).
- Hauck, C., Müller, S. & Schildknecht, H. A germination stimulant for parasitic flowering plants from *Sorghum bicolor*, a genuine host plant. *J. Plant Physiol.* **139**, 474–478 (1992).
- Müller, S., Hauck, C. & Schildknecht, H. Germination stimulants produced by *Vigna unguiculata* Walp cv Saunders Upright. *J. Plant Growth Regul.* **11**, 77–84 (1992).
- Siame, B. A., Weerasuriya, Y., Wood, K., Ejeta, G. & Butler, L. Isolation of strigol, a germination stimulant for *Striga asiatica*, from host plants. *J. Agric. Food Chem.* **41**, 1486–1491 (1993).
- Yokota, T., Sakai, H., Okuno, K., Yoneyama, K. & Takeuchi, Y. Alethrol and orobanchol, germination stimulants for *Orobancha minor*, from its host red clover. *Phytochemistry* **49**, 1967–1973 (1998).
- Johnson, A. W. *et al.* The preparation of synthetic analogues of strigol. *J. Chem. Soc. Perkin Trans. I* **1981**, 1734–1743 (1981).
- Frischmuth, K. *et al.* Routes to derivatives of strigol (the witchweed germination factor) modified in the 5-position. *Tetrahedron* **47**, 9793–9806 (1991).
- Bergmann, C. *et al.* Stimulation of *Orobancha crenata* seed germination by (+)-strigol and structural analogues dependence on constitution and configuration of the germination stimulants. *J. Plant Physiol.* **142**, 338–342 (1993).
- Sugimoto, Y., Wigchert, S. C. M., Thuring, J. W. J. F. & Zwanenburg, B. Synthesis of all eight stereoisomers of the germination stimulant sorgolactone. *J. Org. Chem.* **63**, 1259–1267 (1998).
- Nakano, S., Todoroki, Y., Hirai, N. & Ohigashi, H. Synthesis and biological activity of 7'-, 8'-, and 9'-alkyl analogues of abscisic acid. *Biosci. Biotechnol. Biochem.* **59**, 1699–1706 (1995).
- Brooks, D. W., Bevinakatti, H. S. & Powell, D. R. The absolute structure of (+)-strigol. *J. Org. Chem.* **50**, 3779–3781 (1985).
- Yasuda, N., Sugimoto, Y., Kato, M., Inanaga, S. & Yoneyama, K. (+)-Strigol, a witchweed seed germination stimulant, from *Menispermum dauricum* root culture. *Phytochemistry* **62**, 1115–1119 (2003).
- Mangnus, E. M. & Zwanenburg, B. Tentative molecular mechanisms for germination stimulation of *Striga* and *Orobancha* seeds by strigol and its synthetic analogues. *J. Agric. Food Chem.* **40**, 1066–1070 (1992).
- Westwood, J. H. Characterization of the *Orobancha-Arabidopsis* system for studying parasite-host interactions. *Weed Sci.* **48**, 742–748 (2000).
- Sato, D. *et al.* Analysis of strigolactones, germination stimulants for *Striga* and *Orobancha*, by high-performance liquid chromatography/tandem mass spectrometry. *J. Agric. Food Chem.* **51**, 1162–1168 (2003).
- Yoneyama, K., Takeuchi, Y. & Yokota, T. Production of clover broomrape seed germination stimulants by red clover requires nitrate but is inhibited by phosphate and ammonium. *Physiol. Plant.* **112**, 25–30 (2001).
- Parniske, M. Molecular genetics of the arbuscular mycorrhizal symbiosis. *Curr. Opin. Plant Biol.* **7**, 414–421 (2004).
- Bécard, G. & Fortin, J. A. Early events of vesicular-arbuscular mycorrhiza formation on Ri T-DNA transformed roots. *New Phytol.* **108**, 211–218 (1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Kawaguchi for discussions and critical reading of the manuscript, Y. Sugimoto for providing (±)-sorgolactone, GR24 and (+)-strigol, and K. Yoneyama for discussions and critical reading of the manuscript. This work was supported by Core Research for Evolutional Science and Technology (CREST), Japan Science and Technology Agency, and a Grant-in-Aid for Scientific Research from the Ministry of Education, Science, Sports and Culture of Japan.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to K.A. (akiyama@biochem.osakafu-u.ac.jp).

LETTERS

A microRNA polycistron as a potential human oncogene

Lin He^{1*}, J. Michael Thomson^{2*}, Michael T. Hemann¹, Eva Hernando-Monge⁴, David Mu¹, Summer Goodson², Scott Powers¹, Carlos Cordon-Cardo⁴, Scott W. Lowe¹, Gregory J. Hannon¹ & Scott M. Hammond^{2,3}

To date, more than 200 microRNAs have been described in humans; however, the precise functions of these regulatory, non-coding RNAs remains largely obscure. One cluster of microRNAs, the *mir-17-92* polycistron, is located in a region of DNA that is amplified in human B-cell lymphomas¹. Here we compared B-cell lymphoma samples and cell lines to normal tissues, and found that the levels of the primary or mature microRNAs derived from the *mir-17-92* locus are often substantially increased in these cancers. Enforced expression of the *mir-17-92* cluster acted with *c-myc* expression to accelerate tumour development in a mouse B-cell lymphoma model. Tumours derived from haematopoietic stem cells expressing a subset of the *mir-17-92* cluster and *c-myc* could be distinguished by an absence of apoptosis that was otherwise prevalent in *c-myc*-induced lymphomas. Together, these studies indicate that non-coding RNAs, specifically microRNAs, can modulate tumour formation, and implicate the *mir-17-92* cluster as a potential human oncogene.

MicroRNAs (miRNAs) have emerged relatively recently as a new class of small, non-coding RNAs that regulate gene expression. Nascent primary miRNA transcripts (pri-miRNAs) are processed sequentially by two RNase III enzymes, Drosha and Dicer^{2,3}, to yield mature miRNAs, ranging from 18 to 24 nucleotides (nt) in length. miRNAs are incorporated into the RNA interference (RNAi) effector complex, RISC, and target specific messenger RNAs for translational repression or mRNA cleavage⁴⁻⁶. Although hundreds of miRNAs have been cloned and/or predicted, only a handful have been functionally characterized. For example, *lin-4* and *let-7* regulate the timing of larval development in *C. elegans*^{7,8}. Left/right asymmetric gene expression in *C. elegans* chemosensory neurons is controlled by *lxy-6* and *miR-273* (refs 9, 10). Bantam stimulates cell growth and prevents apoptosis in *Drosophila*¹¹, and *miR-181* potentiates B-cell differentiation in mammals¹². These findings, in combination with computational target predictions, are consistent with miRNAs regulating a broad spectrum of physiological and developmental processes.

Microarray-based expression studies have indicated specific alterations in human miRNA expression profiles that correlate with particular tumour phenotypes (J.M.T. and S.M.H., unpublished data). Among those that show altered expression, the *mir-17-92* cistron is located at 13q31, a genomic locus that is amplified in cases of diffuse large B-cell lymphoma, follicular lymphoma, mantle cell lymphoma, primary cutaneous B-cell lymphoma and several other tumour types^{1,13}. There are only two annotated genes in the epicentre of this amplicon, *c13orf25* and *GPC5*. Previous studies have shown that *c13orf25* is the only one of the two genes for which increased expression correlates with the presence of the amplicon¹. Therefore,

c13orf25 had been implicated as a target of the 13q31 amplicon¹. It is unlikely that *c13orf25* actually encodes a protein, as predicted open reading frames (ORFs) encode only short peptides (<70 amino acids), which are not conserved in closely related species. Instead, the *c13orf25* transcript appears to be the functional precursor of a series of seven microRNAs: *miR-17-5p*, *miR-17-3p*, *miR-18*, *miR-19a*, *miR-20*, *miR-19b-1* and *miR-92-1* (Fig. 1a). Additionally, this cluster is related to the homologous *mir-106a-92* cluster on chromosome X and the *mir-106b-25* cluster on chromosome 7 (ref. 14; Fig. 1a). Alignment of the human *c13orf25* locus and its murine orthologue revealed extensive sequence conservation only within the *mir-17-92* polycistron and its immediate flanking sequence. Several of the expressed-sequence-tags (ESTs) derived from *c13orf25* and its mouse orthologue terminate at the 3' end of *mir-17-92* cluster, consistent with the presence of a Drosha processing site at this location (Fig. 1b).

A principal consequence of 13q31-q32 amplification could be an increase in the abundance of mature miRNA species from the *mir-17-92* cluster. We acquired four cell lines previously described as carrying amplifications in the 13q31-q32 region¹ and confirmed the gene dosage increase at the *c13orf25* locus in three of those cell lines using quantitative polymerase chain reaction (PCR) analysis. The abundance of 191 mature miRNAs was assessed in these four cell lines and compared to normal B-cells, and to five leukaemia and lymphoma cell lines lacking the amplicon (Fig. 1c and Supplementary Fig. 1). Using a significance analysis of microarrays (SAM) analysis¹⁵, we identified six miRNAs for which high-level expression correlated with increased gene dosage of *c13orf25* (see Supplementary Table 1). Five were from the *mir-17-92* cistron, and the sixth, *miR-106a*, was identified as a probable result of cross-hybridization to *miR-17-5p*, from which it differs at only two terminal nucleotides (Fig. 1c). This hypothesis is supported by the observation that the *mir-106a-92* locus does not show copy number alterations in these cell lines (not shown). In each cell line, expression levels correlated with the copy number of the *mir-17-92* locus (Fig. 1c, lower panel).

We also examined the expression of pri-*mir-17-92* in a series of human tumour samples comprising both lymphomas and colorectal carcinomas. Of 46 lymphoma samples, including 13 diffuse large B-cell lymphomas and 6 follicular lymphomas, we saw significant (greater than fivefold) overexpression of pri-*mir-17-92* in 65% of the samples. Considering all of the B-cell lymphoma samples analysed, the average increase in pri-miRNA expression was about tenfold (Fig. 1d). In contrast, colorectal carcinomas rarely showed overexpression of the pri-miRNA. Increases in expression from this locus were less common (15% of samples showed greater than fivefold

¹Cold Spring Harbor Laboratory, Watson School of Biological Sciences, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA. ²Department of Cell and Developmental Biology and ³Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, North Carolina 27599, USA. ⁴Memorial Sloan-Kettering Cancer Center, Division of Molecular Pathology, Memorial Sloan-Kettering Cancer Center, 1275 York Ave, New York, New York 10021, USA.

*These authors contributed equally to this work.

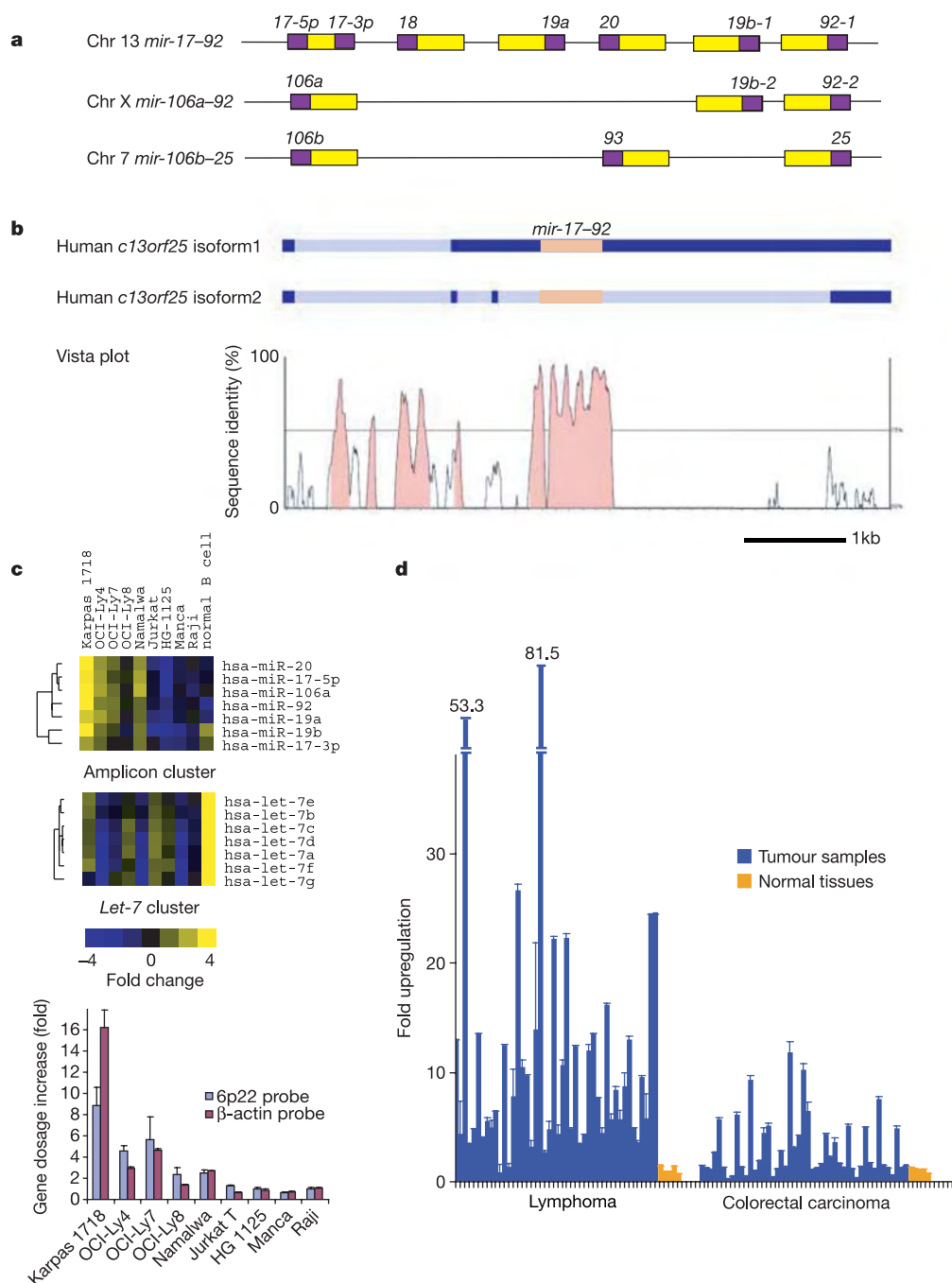


Figure 1 | The *mir-17-92* cluster shows increased expression in B-cell lymphoma samples and cell lines. **a**, Genomic organization of three polycistronic miRNA clusters is shown. There are five paralogous groups located in three homologous clusters (*mir-17-92*, *mir-106a-92* and *mir-106b-25*) with a conserved order: *miR-17-5p/miR-106a/miR-106b*; *miR-18*; *miR-19a/miR-19b-1/miR-19b-2*; *miR-20/miR-93*; and *miR-92-1/miR-92-2/miR-25* (yellow boxes, pre-miRNAs; purple boxes, mature miRNAs). **b**, The level of conservation between human and mouse homologues is represented using an mVista plot²⁸. Two alternative isoforms have been detected for the human gene, and these are shown schematically¹ (dark blue, exons; light blue, introns; orange, the *mir-17-92* cluster). **c**, MicroRNA expression levels in cell lines carrying the 13q31-q32 amplicon (including Karpas 1718, OCI-Ly4 and OCI-Ly7) were compared to those in leukaemia and lymphoma cell lines lacking this genetic lesion, and to normal B-cells isolated from cortical blood (top panel). We included in this analysis

the OCI-Ly8 cell line, which has previously been identified as a cell line carrying the 13q31-32 amplicon, but showed no gene dosage increase at the *c13orf25* locus in our study. Normalized one-channel measurements for 191 human miRNAs were hierarchically clustered for all miRNA genes represented on the array. An excerpt of the data is shown, and the full cluster analysis is presented in Supplementary Fig. 1. The expression map node that correlates with the amplification is shown. The *let-7* cluster node is also shown for comparison (middle panel). In the cell lines examined, expression levels of the mature microRNAs from the *mir-17-92* polycistron correlate with the copy number at the *mir-17-92* locus (bottom panel). **d**, The level of *mir-17-92* pri-miRNA was determined by real-time quantitative PCR in 46 lymphomas and 47 colorectal carcinomas, and compared to levels found in corresponding normal tissues from five individuals. In **c** and **d**, error bars indicate standard deviation (s.d.).

upregulation), and the degree of overexpression was substantially lower (Fig. 1d).

Considered together, our data prompted the hypothesis that *mir-17-92* might contribute to tumour development. To test this idea directly, we used a mouse model of human B-cell lymphoma. Transgenic animals carrying a *c-myc* oncogene, driven by the immunoglobulin heavy-chain enhancer (*E μ*), develop B-cell lymphomas by 4–6 months of age¹⁶. Similarly, haematopoietic stem cells (HSCs) derived from fetal livers of *E μ -myc* transgenic mice generate B-cell lymphomas with comparable latency when transplanted into lethally irradiated recipients^{17–20} (Fig. 2a). We therefore infected *E μ -myc*+ HSCs with a murine stem cell virus (MSCV) retrovirus that directs expression of a truncated cluster comprising *mir-17-19b-1* (hereafter *mir-17-19b*), the vertebrate-specific portion of the *mir-17-92* miRNA cistron (Fig. 1a). This virus also contained a green fluorescent protein (GFP) transgene, allowing us to follow infected stem cells *in vitro* and *in vivo* (Fig. 2a). Mice reconstituted with *E μ -myc*+ HSCs carrying a control MSCV vector developed lymphomas after the expected latency (3–6 months), with incomplete penetrance (Fig. 2b). Similarly, we examined >40 animals reconstituted with *E μ -myc*+ HSCs expressing subsets of 96 different, single microRNAs (see Supplementary Table 2). Although we did not

confirm miRNA overexpression for every case, we did not observe any significantly accelerated onset of disease. In contrast, 100% of the animals co-expressing the *mir-17-19b* polycistron and *c-myc* developed leukaemias at an average of 51 days following transplantation (s.d. ± 13 days, $P < 0.0001$ compared with MSCV controls using the logrank test), and died of B-cell lymphomas at an average age of 65 days (s.d. ± 13 days, $P < 0.0001$ compared to MSCV controls; Fig. 2b). In all but one case, primary lymphomas could be visualized by virtue of the linked GFP marker (Fig. 2c and Table 1). The mature miRNAs from the *mir-17-19b* cluster show high-level expression in these tumours, compared with miRNAs from the paralogous *mir-106a-92* locus, and also have similar *mir-17-19b* expression levels compared to the Karpas 1718 lymphoma cell line, which has increased *c13orf25* gene dosage (see Supplementary Fig. 2).

The full *mir-17-92* cistron was also tested in a small cohort of animals. Although similar results were obtained compared to those animals reconstituted with HSCs expressing *mir-17-19b*, studies in cell lines indicated that the construct used to express the entire cluster gave lower levels of mature miRNAs. We therefore focused most of our study on the truncated *mir-17-19b* cluster. In these ongoing studies, we have yet to find any individual miRNA from the

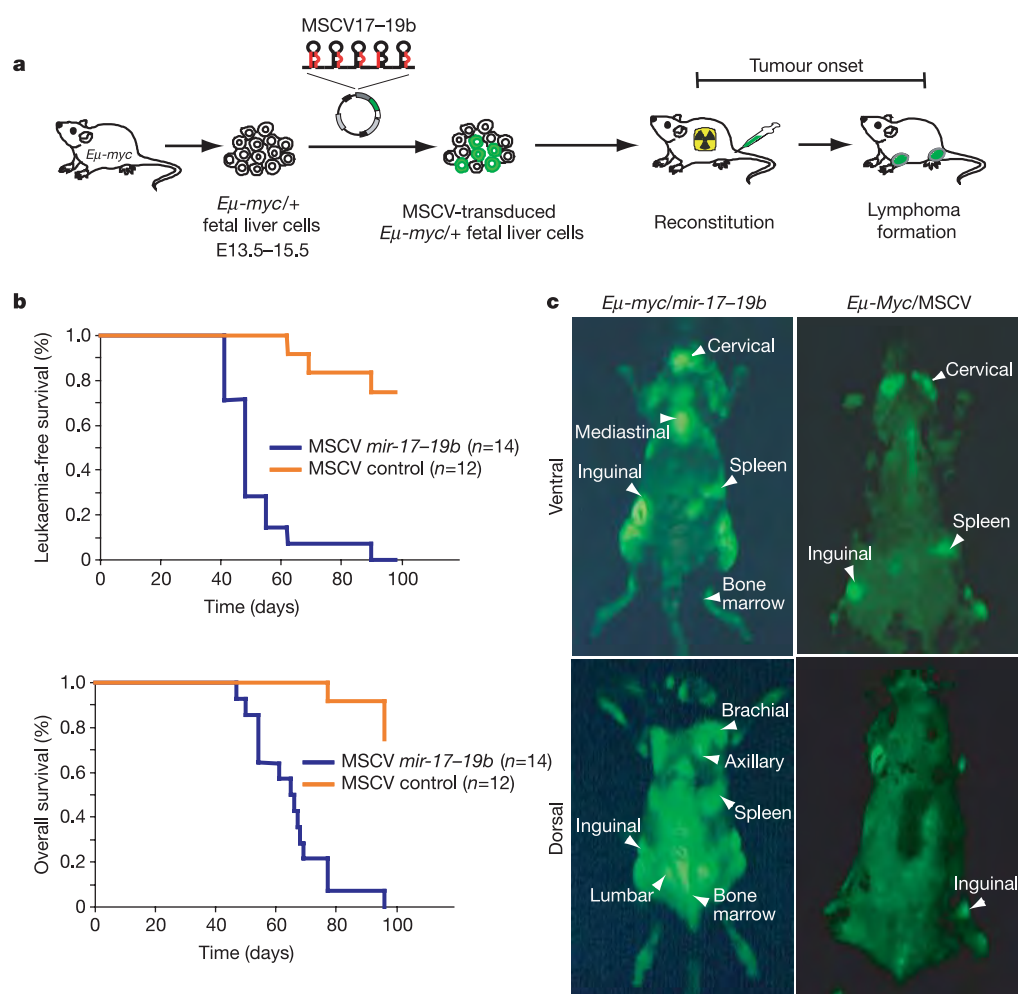


Figure 2 | Overexpression of the *mir-17-19b* cluster accelerates *c-myc*-induced lymphomagenesis in mice. **a**, Schematic representation of the adoptive transfer protocol using *E μ -myc* HSCs. **b**, Mice reconstituted with HSCs expressing *mir-17-19b* in an MSCV retroviral vector (MSCV *mir-17-19b*) or infected with a control MSCV virus were monitored by blood smear analysis starting 5 weeks after transplantation. The Kaplan-

Meier curves represent the percentage of leukaemia-free survival or overall survival. **c**, External GFP imaging of tumour-bearing mice with *E μ -myc/mir-17-19b* or *E μ -myc/MSCV* shows the overall distribution of tumour cells. *E μ -myc/mir-17-19b* tumours show a more disseminated phenotype compared with control tumours. These animals are representative of their genotype.

Table 1 | Phenotypic analysis of a subset of *Eμ-myc/mir-17-19b* tumours

Animal	GFP	Immunophenotyping	Cell type	Ki67 staining	Apoptosis*	Pathological features
1	+	B220 ⁺ , Thy1.2 [−] , IgM [−] , CD19 ⁺ , CD4 [−] , CD8 [−]	Pre-B cell	++(70–80%)	Low	Tumour cells invaded liver and spleen; mild infiltrations observed in lung and kidney; spleen enlarged; hindlimb paralysis
2	+	B220 ⁺ , Thy1.2 ^{low} , IgM [−] , CD19 ⁺ , CD4 [−] , CD8 [−]	Pre-B cell	++(80–90%)	Low	Tumour cells invaded liver, lung and spleen; spleen enlarged
3	+	B220 ⁺ , Thy1.2 [−] , IgM [−] , CD19 ⁺ , CD4 [−] , CD8 [−]	Pre-B cell	++(80–90%)	Low	Tumour cells invaded liver, lung and spleen; spleen enlarged
4	+	B220 ⁺ , Thy1.2 ^{low} , IgM [−] , CD19 ⁺ , CD4 [−] , CD8 [−]	Pre-B cell	++(70–80%)	Low	Tumour cells invaded liver and spleen; spleen enlarged
5	+	B220 ⁺ , Thy1.2 [−] , IgM [−] , CD19 ⁺ , CD4 [−] , CD8 [−]	Pre-B cell	++(80–90%)	Slightly less than control	Highly disseminated lymphoma; tumour cells invaded liver, spleen, lung and kidney; spleen enlarged
6	+	B220 ⁺ , Thy1.2 [−] , IgM [−] , CD19 ⁺ , CD4 [−] , CD8 [−]	Pre-B cell	++(80–90%)	Low	Highly disseminated lymphoma; tumour cells invaded liver, spleen, lung and kidney; spleen enlarged
7	+	B220 ⁺ , Thy1.2 [−] , IgM [−] , CD19 ⁺ , CD4 [−] , CD8 [−]	Pre-B cell	++(70–80%)	Low	Spleen enlarged; mild infiltration of tumour cells into liver only
8	+	B220 ⁺ , Thy1.2 ^{low} , IgM [−] , CD19 ⁺ , CD4 [−] , CD8 [−] and B220 ⁺ , Thy1.2 ^{low} , IgM [−] , CD19 ⁺ , CD4 [−] , CD8 [−] †	Pre-B cell and mature B cell	++(70–80%)	Low	Highly disseminated lymphoma; tumour cells invaded liver, spleen, lung and kidney; enlarged spleen
9	−	ND	ND	++(80–90%)	Low	Tumour cells invaded liver, lung and spleen; enlarged spleen

*Levels of apoptosis in *Eμ-myc/mir-17-19b* lymphomas are compared to control *Eμ-myc/MSCV* lymphomas on the basis of haematoxylin and eosin staining, and TUNEL staining.

†The two clones of tumour cells with different cell type specificity might reflect independent transformation events or maturation of a single primary clone.

ND, not determined.

mir-17-19b cluster that can accelerate tumour formation to the extent seen with the intact polycistron (not shown).

The *Eμ-myc/mir-17-19b* lymphomas are true malignancies rather than hyperplasias, because primary tumour cells, when transplanted into C57B6/J recipients, induce B-cell lymphomas in 2–3 weeks that result in lethality after 4–5 weeks (data not shown). The secondary tumours show pathological features indistinguishable from the original tumours, and retain tumorigenic potential after two additional rounds of serial transplantation (data not shown). Therefore, a microRNA cluster can accelerate *Eμ-myc* induced tumorigenesis in mice.

The pathological hallmarks of *Eμ-myc/mir-17-19b* mosaic animals included massive enlargement of lymph nodes, splenic hyperplasia, infiltration of the thymus by lymphoma cells, and leukaemia (Fig. 2c). Animals with advanced lymphomas showed extramedullary haematopoiesis due to functional failure of the bone marrow. Furthermore, 6 out of 14 animals showed hind limb paralysis, associated with substantial tumour growth at the lumbar node. Tumours resulting from combined *c-myc* and *mir-17-19b* expression consistently invaded visceral organs outside the lymphoid compartment, including liver, lung and occasionally kidney (Figs 2c, 3b and Table 1). Additionally, *Eμ-myc/mir-17-19b* lymphomas show a high mitotic index without extensive apoptosis (Fig. 3a). This contrasts with the *Eμ-myc/MSCV* tumours lacking the microRNA cluster, which show a high degree of apoptosis (Fig. 3a). These findings indicate that cooperation between *Eμ-myc* and *mir-17-19b* gives rise to highly malignant, disseminated lymphomas capable of evading normal apoptotic responses to inappropriate proliferation.

Eμ-myc-induced lymphomas originate from the B-lymphoid lineage. However, the developmental characteristics of these tumour cells are not stage-specific, as they can resemble either mature B cells or pre-B cells. To examine the cell lineage of the *Eμ-myc/mir-17-19b* lymphomas, we assessed the expression of cell-surface markers, including the B-cell-specific marker B220 and the T-cell-specific markers CD4 and CD8a. All tumours were found to be of B-cell origin, staining positive for B220 and negative for both CD4 and CD8a (Fig. 3c and Table 1). We next analysed these tumours for CD19 and IgM expression to distinguish pre-B from mature B cells. With one exception, *Eμ-myc/mir-17-19b* lymphomas were derived

purely from the pre-B cell lineage (low to absent Thy1.2 staining, CD19⁺B220⁺IgM[−]) (Table 1), suggesting that overexpression of *mir-17-19b* strongly favours transformation of B-cell progenitors under our experimental conditions.

Exactly how overproduction of *mir-17-19b* might promote oncogenesis is unclear. At a minimum, studies of tumour pathology suggest that increased expression of this cluster mitigates the pro-apoptotic response to elevated *myc* expression *in vivo*. Furthermore, we have previously shown that this miRNA cluster is highly expressed in embryonic stem cells, with expression levels decreasing during embryonic development in mice²¹. It is therefore tempting to speculate that these miRNAs promote 'stem cell' properties or specify characteristics of early developmental lineages. A detailed, mechanistic understanding of how this non-coding RNA cluster acts as an oncogene is at present hampered by the lack of a validated biochemical strategy for identifying miRNA targets.

Previous circumstantial evidence has indicated the potential involvement of a number of miRNAs in tumorigenesis. Although miRNAs only represent 1% of the mammalian genome, more than 50% of miRNA genes are located within regions associated with amplification, deletion and translocation in cancer²². Expression studies of various tumour types have also revealed specific alterations in miRNA profiles^{22–25}. For example, *mir-15* and *mir-16* are frequently deleted and/or downregulated in B-cell chronic lymphocytic leukaemia²⁶, miR-143 and miR-145 show decreased abundance in colorectal neoplasia²⁵, and miR-155 and its non-coding RNA host gene, *BIC*, are upregulated 100-fold in Burkitt's lymphoma patients²⁴. Here, we have shown that one miRNA polycistron is not only the subject of tumour-specific amplification, but that it is also overexpressed in tumours and tumour cell lines, and can act as an oncogene *in vivo*. Our results indicate that non-coding RNAs might act as integral parts of the molecular architecture of oncogene and tumour suppressor networks. Such oncogenic microRNAs might be designated 'oncomiRs', with *mir-17-92* being oncomiR-1.

METHODS

miRNA expression profiling. Five micrograms of total RNA was labelled with RNA ligase and a Cy3-conjugated dinucleotide, and hybridized to custom oligonucleotide microarrays as described in ref. 21. Cy3 median intensity values

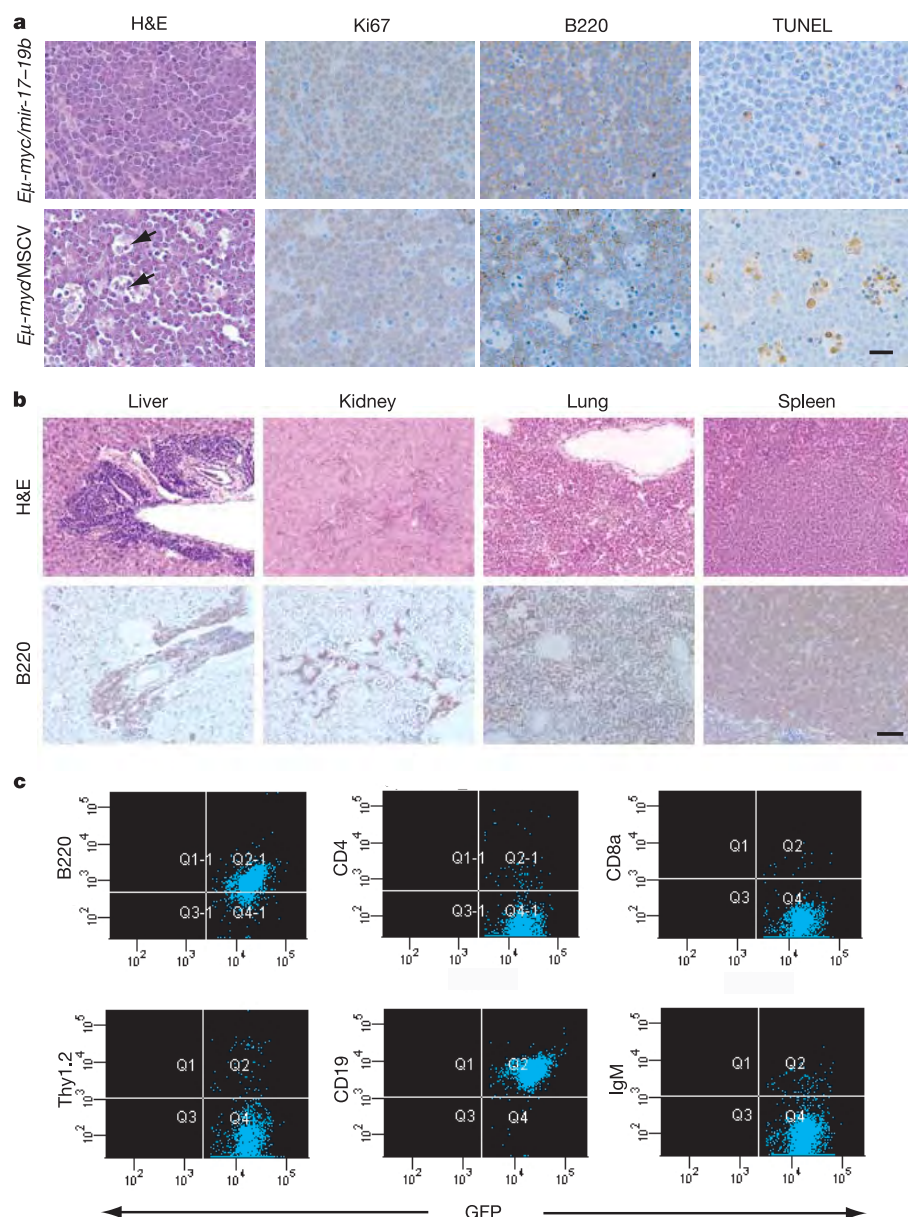


Figure 3 | Pathological and immunological analysis of lymphomas produced by cooperation between *mir-17-19b* and *c-myc*. **a**, Haematoxylin and eosin (H&E), Ki67, B220 and TUNEL staining of *Eμ-myc/mir-17-19b* lymph node tumours. The 'starry sky' morphology is a hallmark of cell clusters undergoing apoptosis (black arrows). Scale bar, 10 μ m. **b**, Invasion of visceral organs (liver, spleen, lung and kidney) by *Eμ-myc/mir-17-19b* tumour cells, shown by H&E and B220 staining. Invasion was observed both

perivascularly and parenchymally in liver. Scale bar, 50 μ m.

c, Immunophenotyping of *Eμ-myc/mir-17-19b* lymphomas. Tumour cells stained positively for the B-cell-specific marker B220, but not for the T-cell-specific markers CD4, CD8a and Thy1.2. Tumour cells bore cellular characteristics of pre-B cells, staining positively for CD19 but not for IgM, a marker of mature B-cells.

were filtered to remove data points that did not exceed background levels by twofold. Values were log₂-transformed and median-centred by array. Clustering was performed with Cluster (Stanford University), using values that were median-centred by gene. Dendrograms and expression maps were generated using Treeview (Stanford).

Cell lines. The measurement of miRNA abundance was carried out using the following human cell lines: Karpas 1718 (derived from splenic lymphoma with villous lymphocytes, provided by A. Karpas), OCI-Ly4, OCI-Ly7 and OCI-Ly8 (derived from diffuse large B-cell lymphoma, provided by R. Dalla-Favera). The cell lines lacking the 13q31-q32 amplicon were Raji (B-cell, derived from Burkitt's lymphoma, ATCC); Namalwa (B-cell, derived from Burkitt's lymphoma, ATCC); HG 1125 (EBV-transformed human lymphoblastoid, provided by B. Stillman); Manca (lymphoblast-like, derived from chronic myelogenous leukaemia); Jurkat; proliferating B-cells (splenic B-cells isolated from a C57B6/Ly5.2 mouse and stimulated to proliferate in culture with lipopolysaccharide,

provided by I. Weissman); and normal B cells (derived from cord blood, Cambrex).

PCR and copy number analysis. Tumour samples were obtained from the Cooperative Human Tissue Network (<http://www-chn.ims.nci.nih.gov>). Corresponding normal tissue RNA from five individuals was purchased from Biochain Institute Inc. For fluorogenic real-time PCR, primers that amplify *mir-17-92* pri-miRNA and β -actin mRNA (control), and probes were designed using Primer Express software (V.2): *mir-17-92* forward primer, 3'-CAGTAAA GGTAAAGGAGAGCTCAATCTG-5'; reverse primer, 3'-CATACAACCACTAA GCTAAAGAATAATCTGA-5'; *mir-17-92* probe, (6-FAM)-TGGAATAAGATC ATCATGCCCACTTGAGAC-(TAMRA); β -actin forward primer, 3'-GCAAAG ACCTGTACGCCAACA-5'; reverse primer, 3'-TGCATCCTGTGGCAATG-5'; β -actin probe, (6-FAM)-TGGCGGCACCACCATGTACC-(TAMRA). The ratios of RNA species detected by *mir-17-92* primers and β -actin primers in each RNA sample were determined in triplicate by reverse-transcription, quantitative PCR

using an ABI 7900HT Taqman sequence detector, following the 'standard curve' method. All values were subsequently normalized to the averaged ratio of the five corresponding normal samples. For DNA copy number determination using the ABI 7900HT sequence detector, we performed quantitative PCR analysis using the same *mir-17-92* primer set described above, and normalized the data against a reference probe corresponding to chromosomal region 6p22 (forward primer, 3'-GGTCTCTATTGCACTTGGCTGAT-5'; reverse primer, 3'-TTTCA TTGTTGACCAAGCTAGACA-5'; probe, (6-FAM)-TAGGGCATACTGCCTG CATATTTCTTGCT-(TAMRA)) or a β -actin probe. The reported values represent the ratios of DNA copy number at the *mir-17-92* locus to the normal reference probe.

Adoptive transfer of *Eμ-myc* HSCs. Fetal liver-derived HSCs were isolated from *Eμ-myc/+* mouse embryos at embryonic day (E)13.5–E15.5, and were transduced with MSCV alone or MSCV expressing the *mir-17-19b* cluster. To exclude the possibility that the observed acceleration of lymphomagenesis was due to insertional mutagenesis, independent experiments were carried out using fetal liver cells isolated from eight *Eμ-myc/+* embryos. Cells from each isolation were separately infected with MSCV *mir-17-19b* or control MSCV. The MSCV retroviral vector used in our studies contains a PGK-puromycin-IRES-GFP (PIG) cassette¹⁸. Infection rates, assessed by fluorescence-activated cell sorting (FACS), were typically 40% of bulk fetal liver cells. HSCs infected with MSCV-*mir17-19b*-PIG and MSCV-PIG (control) were subsequently transplanted into 6–8-week-old, lethally irradiated C57BL/6 recipient mice¹⁷. Tumour onset was monitored by blood smear analysis, and tumour samples were either collected into 4% paraformaldehyde for histopathological studies, or prepared as single-cell suspensions for FACS.

We also carried out a screen of 96 miRNAs, to look for miRNA(s) that accelerate Myc-induced lymphomagenesis. In this experiment, each pre-miRNA (and ~50 bp of flanking sequence 5' and 3' of the pre-miRNA) was cloned downstream of the cytomegalovirus (CMV) promoter in a MSCV vector containing SV40–GFP. Eight individual MSCV constructs, each overexpressing a specific miRNA, were pooled at equal concentrations. Twelve pools of DNA were used individually to produce virus to infect *Eμ-myc/+* fetal liver cells for adoptive transfer into at least three recipient animals, as described above. Recipient animals were monitored for tumour growth for at least six months. For those that developed lymphoma, tumour cells were prepared from the enlarged lymph nodes and then subjected to FACS analysis for GFP expression. **Histopathology.** Tissue samples were fixed in 4% paraformaldehyde, embedded in paraffin, sectioned into 5-μm slices, and stained with haematoxylin and eosin. For Ki67 detection (rabbit anti-Ki67 antibody, NovoCastra), representative sections were deparaffinized, rehydrated in graded alcohols, and processed using the avidin–biotin immunoperoxidase method. Sections were then subjected to antigen retrieval by microwave oven treatment using standard procedures. Diaminobenzidine was used as the chromogen and haematoxylin as the nuclear counterstain. For B220 immunohistochemistry, a rat anti-mouse CD45R/B220 antibody (clone RA3-6B2, BD Biosciences Pharmingen) was used; pretreatment for antigen retrieval was not required. Analysis of the rate of apoptosis by TUNEL assay was performed according to a published protocol²⁷.

Received 15 February; accepted 16 March 2005.

- Ota, A. *et al.* Identification and characterization of a novel gene, *C13orf25*, as a target for 13q31-q32 amplification in malignant lymphoma. *Cancer Res.* **64**, 3087–3095 (2004).
- Lee, Y. *et al.* The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425**, 415–419 (2003).
- Bernstein, E., Caudy, A. A., Hammond, S. M. & Hannon, G. J. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409**, 363–366 (2001).
- Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297 (2004).
- Ambros, V. The functions of animal microRNAs. *Nature* **431**, 350–355 (2004).
- He, L. & Hannon, G. J. MicroRNAs: small RNAs with a big role in gene regulation. *Nature Rev. Genet.* **5**, 522–531 (2004).
- Ruvkun, G. & Giusto, J. The *Caenorhabditis elegans* heterochronic gene *lin-14* encodes a nuclear protein that forms a temporal developmental switch. *Nature* **338**, 313–319 (1989).
- Ambros, V. A hierarchy of regulatory genes controls a larva-to-adult developmental switch in *C. elegans*. *Cell* **57**, 49–57 (1989).
- Chang, S., Johnston, R. J. Jr, Frokjaer-Jensen, C., Lockery, S. & Hobert, O.

- MicroRNAs act sequentially and asymmetrically to control chemosensory laterality in the nematode. *Nature* **430**, 785–789 (2004).
- Johnston, R. J. & Hobert, O. A microRNA controlling left/right neuronal asymmetry in *Caenorhabditis elegans*. *Nature* **426**, 845–849 (2003).
- Brennecke, J., Hipfner, D. R., Stark, A., Russell, R. B. & Cohen, S. M. *bantam* encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene *hid* in *Drosophila*. *Cell* **113**, 25–36 (2003).
- Chen, C. Z., Li, L., Lodish, H. F. & Bartel, D. P. MicroRNAs modulate hematopoietic lineage differentiation. *Science* **303**, 83–86 (2004).
- Knuutila, S. *et al.* DNA copy number amplifications in human neoplasms: review of comparative genomic hybridization studies. *Am. J. Pathol.* **152**, 1107–1123 (1998).
- Tanzer, A. & Stadler, P. F. Molecular evolution of a microRNA cluster. *J. Mol. Biol.* **339**, 327–335 (2004).
- Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA* **98**, 5116–5121 (2001).
- Adams, J. M. *et al.* The *c-myc* oncogene driven by immunoglobulin enhancers induces lymphoid malignancy in transgenic mice. *Nature* **318**, 533–538 (1985).
- Schmitt, C. A. *et al.* Dissecting p53 tumor suppressor functions *in vivo*. *Cancer Cell* **1**, 289–298 (2002).
- Hemann, M. T. *et al.* An epi-allelic series of p53 hypomorphs created by stable RNAi produces distinct tumor phenotypes *in vivo*. *Nature Genet.* **33**, 396–400 (2003).
- Hemann, M. T. *et al.* Suppression of tumorigenesis by the p53 target PUMA. *Proc. Natl Acad. Sci. USA* **101**, 9333–9338 (2004).
- Wendel, H. G. *et al.* Survival signalling by Akt and eIF4E in oncogenesis and cancer therapy. *Nature* **428**, 332–337 (2004).
- Thomson, J. M., Parker, J., Perou, C. M. & Hammond, S. M. A custom microarray platform for analysis of microRNA gene expression. *Nature Methods* **1**, 47–53 (2004).
- Calin, G. A. *et al.* Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proc. Natl Acad. Sci. USA* **101**, 2999–3004 (2004).
- Calin, G. A. *et al.* MicroRNA profiling reveals distinct signatures in B cell chronic lymphocytic leukemias. *Proc. Natl Acad. Sci. USA* **101**, 11755–11760 (2004).
- Metzler, M., Wilda, M., Busch, K., Viehmann, S. & Borkhardt, A. High expression of precursor microRNA-155/*BIC* RNA in children with Burkitt lymphoma. *Genes Chromosom. Cancer* **39**, 167–169 (2004).
- Michael, M. Z., O'Connor, S. M., van Holst Pellekaan, N. G., Young, G. P. & James, R. J. Reduced accumulation of specific microRNAs in colorectal neoplasia. *Mol. Cancer Res.* **1**, 882–891 (2003).
- Calin, G. A. *et al.* Frequent deletions and down-regulation of micro-RNA genes *miR15* and *miR16* at 13q14 in chronic lymphocytic leukemia. *Proc. Natl Acad. Sci. USA* **99**, 15524–15529 (2002).
- Di Cristofano, A., De Acetis, M., Koff, A., Cordon-Cardo, C. & Pandolfi, P. P. Pten and p27^{KIP1} cooperate in prostate cancer tumor suppression in the mouse. *Nature Genet.* **27**, 222–224 (2001).
- Mayor, C. *et al.* VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16**, 1046–1047 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank members of the Hannon, Lowe and Hammond laboratories for discussions and input. We also thank Z. Xuan, N. Chen, N. Sheth and R. Sachidanandam for bioinformatic analysis. C. Perou and J. Leib provided advice and support for microarray methodologies, and A. Barnes and B. Boone gave assistance with mouse tissue procurement. We are grateful to H. Wendel, C. Scott, C. Marsden and C. Rosenthal, R. Karni, P. Moody and R. Whitaker, who provided advice and technical support. F. Slack coined the oncomiR nomenclature. L.H. and M.T.H. are Fellows of the Helen Hay Whitney Foundation. S.W.L. and C.C.-C. are supported by a program project grant from the NCI. G.J.H. is supported by an Innovator Award from the US Army Breast Cancer Research Program and by grants from the DOD and NIH. S.M.H. is a General Motors Cancer Research Foundation Scholar, and J.M.T. is a Frederick Gardner Cottrell Postdoctoral Fellow.

Author Information Microarray data have been deposited in NCBI-GEO under accession numbers GSM45026–GSM45065 and GSE-2399. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to G.J.H. (hannon@cshl.edu) or S.M.H. (hammond@med.unc.edu).

LETTERS

MicroRNA expression profiles classify human cancers

Jun Lu^{1,4*}, Gad Getz^{1*}, Eric A. Miska^{2*†}, Ezequiel Alvarez-Saavedra², Justin Lamb¹, David Peck¹, Alejandro Sweet-Cordero^{3,4}, Benjamin L. Ebert^{1,4}, Raymond H. Mak^{1,4}, Adolfo A. Ferrando⁴, James R. Downing⁵, Tyler Jacks^{2,3}, H. Robert Horvitz² & Todd R. Golub^{1,4,6}

Recent work has revealed the existence of a class of small non-coding RNA species, known as microRNAs (miRNAs), which have critical functions across various biological processes^{1,2}. Here we use a new, bead-based flow cytometric miRNA expression profiling method to present a systematic expression analysis of 217 mammalian miRNAs from 334 samples, including multiple human cancers. The miRNA profiles are surprisingly informative, reflecting the developmental lineage and differentiation state of the tumours. We observe a general downregulation of miRNAs in tumours compared with normal tissues. Furthermore, we were able to successfully classify poorly differentiated tumours using miRNA expression profiles, whereas messenger RNA profiles were highly inaccurate when applied to the same samples. These findings highlight the potential of miRNA profiling in cancer diagnosis.

Much progress has been made over the last decade in developing a molecular taxonomy of cancer (see ref. 3). In particular, it has become clear that among the ~22,000 protein-coding transcripts are mRNAs that can be used to classify a wide variety of human cancers⁴. Recently, hundreds of small, non-coding miRNAs have been discovered (see ref. 1). The first identified miRNAs, the products of the *C. elegans* genes *lin-4* and *let-7*, have important roles in controlling developmental timing and probably act by regulating mRNA translation^{5–7}. When *lin-4* or *let-7* is inactivated, specific epithelial cells undergo additional cell divisions instead of their normal differentiation. Because abnormal cell proliferation is a hallmark of human cancers, it seems possible that miRNA expression patterns might denote the malignant state. Indeed, altered expression of a few miRNAs has been found in some tumour types^{8–11}. However, the potential for miRNA expression to inform cancer diagnosis has not been systematically explored.

To determine the expression pattern of all known miRNAs, we first needed to develop an accurate and inexpensive profiling method. This goal is challenging, because of the short size of miRNAs (about 21 nucleotides) and the sequence similarity between miRNA family members. Glass-slide microarrays have been used for miRNA profiling^{12–18}, but cross-hybridization of related miRNAs has been problematic. We therefore developed a bead-based profiling method. Oligonucleotide-capture probes complementary to miRNAs of interest were coupled to carboxylated 5-micron polystyrene beads impregnated with variable mixtures of two fluorescent dyes (that can yield up to 100 colours), each representing a single miRNA. Following adaptor ligations which use both the

5'-phosphate and the 3'-hydroxyl groups of miRNAs¹³, reverse-transcribed miRNAs were (1) amplified by polymerase chain reaction (PCR) using a common biotinylated primer, (2) hybridized to the capture beads, and (3) stained with streptavidin-phycoerythrin. The beads were then analysed using a flow cytometer capable of measuring bead colour (denoting miRNA identity) and phycoerythrin intensity (denoting miRNA abundance) (see Supplementary Fig. 1).

Bead-based hybridization has the theoretical advantage that it might more closely approximate hybridization in solution, and as such, we might expect the specificity to be superior to glass microarray hybridization. Indeed, a spiking experiment involving 11 related sequences showed increased specificity of bead-based detection compared with microarray-based detection, even for single base-pair mismatches (Fig. 1a, b). In addition, the bead method showed linear detection over a hundred-fold range of expression (see Supplementary Information). The expression patterns of eight miRNAs in seven different cell lines were validated by northern blotting. In all cases, bead-based detection paralleled the data from northern blotting (Fig. 1c). These results demonstrate that bead-based miRNA detection is feasible, and has the attractive properties of improved accuracy, high speed and low cost. The bead-based detection platform also provides flexibility, in that additional miRNA capture beads can be added to the mixture, allowing detection of newly discovered miRNAs.

We then set out to determine the expression pattern of all known miRNAs across a large panel of samples representing diverse human tissues and tumour types. Although miRNA expression has been explored in small sets of tissues^{12,14–18} or isolated cell types (for example, chronic lymphocytic leukaemia¹⁹), the extent of differential miRNA expression across cancers has not been determined. Indeed, we might not expect that miRNA expression patterns could be informative with respect to cancer diagnosis, because of the relatively small number of miRNAs encoded in the genome. However, we observed differential expression of nearly all miRNAs across cancer types (Fig. 2a). Moreover, hierarchical clustering of the samples using miRNA profiles paralleled the developmental origins of the tissues. For example, samples of epithelial origin were located on a single branch of the dendrogram, whereas the other major branch was predominantly populated with samples of haematopoietic malignancies.

Furthermore, the miRNAs partitioned tumours within a single lineage. For example, we examined the miRNA profiles of 73 bone

¹Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02141, USA. ²Howard Hughes Medical Institute, Department of Biology, Massachusetts Institute of Technology, and ³MIT Center for Cancer Research, Cambridge, Massachusetts 02139, USA. ⁴Department of Pediatric Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, Massachusetts 02115, USA. ⁵Department of Pathology, St. Jude Children's Research Hospital, Memphis, Tennessee 38105, USA. ⁶Howard Hughes Medical Institute, Harvard Medical School, Boston, Massachusetts 02115, USA. [†]Present address: Wellcome Trust/Cancer Research UK, Gurdon Institute, University of Cambridge, Cambridge CB2 1QN, UK.

*These authors contributed equally to this work.

marrow samples obtained from patients with acute lymphoblastic leukaemia (ALL). As shown in Fig. 2b, hierarchical clustering revealed non-random partitioning of the samples into three major branches: one containing all five t(9;22) *BCR/ABL*-positive samples and 10 out of 11 t(12;21) *TEL/AML1* samples; a second branch containing 15 out of 19 T-cell ALL samples; and a third branch containing all but one of the samples with an *MLL* gene rearrangement. These experiments demonstrate that even within a single developmental lineage, distinct patterns of miRNA expression can be observed that reflect mechanisms of transformation, and further support the idea that miRNA expression patterns encode the developmental history of human cancers.

Among the epithelial samples, those of the gastrointestinal tract were of particular interest. Samples from colon, liver, pancreas and stomach all clustered together (Fig. 2a), reflecting their common derivation from tissues of embryonic endoderm. We suggest that sample clustering in miRNA space is predominantly driven by developmental history. In contrast, when the same samples were profiled in the space of ~16,000 mRNAs, the coherence of gut-derived samples was not observed in hierarchical clustering (Fig. 2c). This observation might result from the large amount of noise and unrelated signals that are embedded in the high-dimensional mRNA data. Whether or not the miRNAs that are highly expressed in the gut-associated cluster (miR-192, miR-194 and miR-215) have a functional role in the specification of gut development or gut-derived tumours remains to be investigated.

Having determined that miRNA expression distinguishes tumours of different developmental origin, we next asked whether miRNAs could be used to distinguish tumours from normal tissues. We have previously reported that there are no robust mRNA markers that show consistent differential expression between tumours and normal

tissues of different lineages⁴. It was therefore striking to observe that despite the fact that some miRNA expression levels were upregulated or unchanged, most of the miRNAs (129 out of 217, $P < 0.05$ after correction for multiple hypothesis testing) had lower expression levels in tumours compared with normal tissues, irrespective of cell type (Fig. 3a). Cancer cell lines also showed low miRNA expression relative to normal tissues (see Supplementary Fig. 4).

To exclude any possibility that this differential miRNA expression might be related to differences in the collection of tumour samples versus normal samples, we studied a mouse model of *K-Ras*-induced lung cancer²⁰. We isolated miRNAs from normal lung or lung adenocarcinomas from different individual mice, thereby precluding any differences in collection procedure. Owing to miRNA sequence conservation between human and mouse, the same miRNA capture beads could be used to profile the murine samples. As shown in Fig. 3b, the same distinction between normal and tumour samples is observed in mouse. Accordingly, a tumour/normal classifier constructed with human samples had 100% accuracy when tested in the mouse. Taken together, these studies indicate that miRNAs are unexpectedly rich in information content with respect to cancer.

Our observation that miRNA expression seems globally higher in normal tissues compared with tumours led us to the hypothesis that global miRNA expression reflects the state of cellular differentiation. To test this hypothesis, we explored an experimental model in which we treated the myeloid leukaemia cell line HL-60 with *all-trans* retinoic acid, a potent inducer of neutrophilic differentiation²¹. As predicted, miRNA profiling revealed the induction of many miRNAs coincident with differentiation (Fig. 3c). In primary human haematopoietic progenitor cells undergoing erythroid differentiation *in vitro*, we observed a similar increase in miRNA expression at a stage in differentiation when the cells continued to proliferate (see

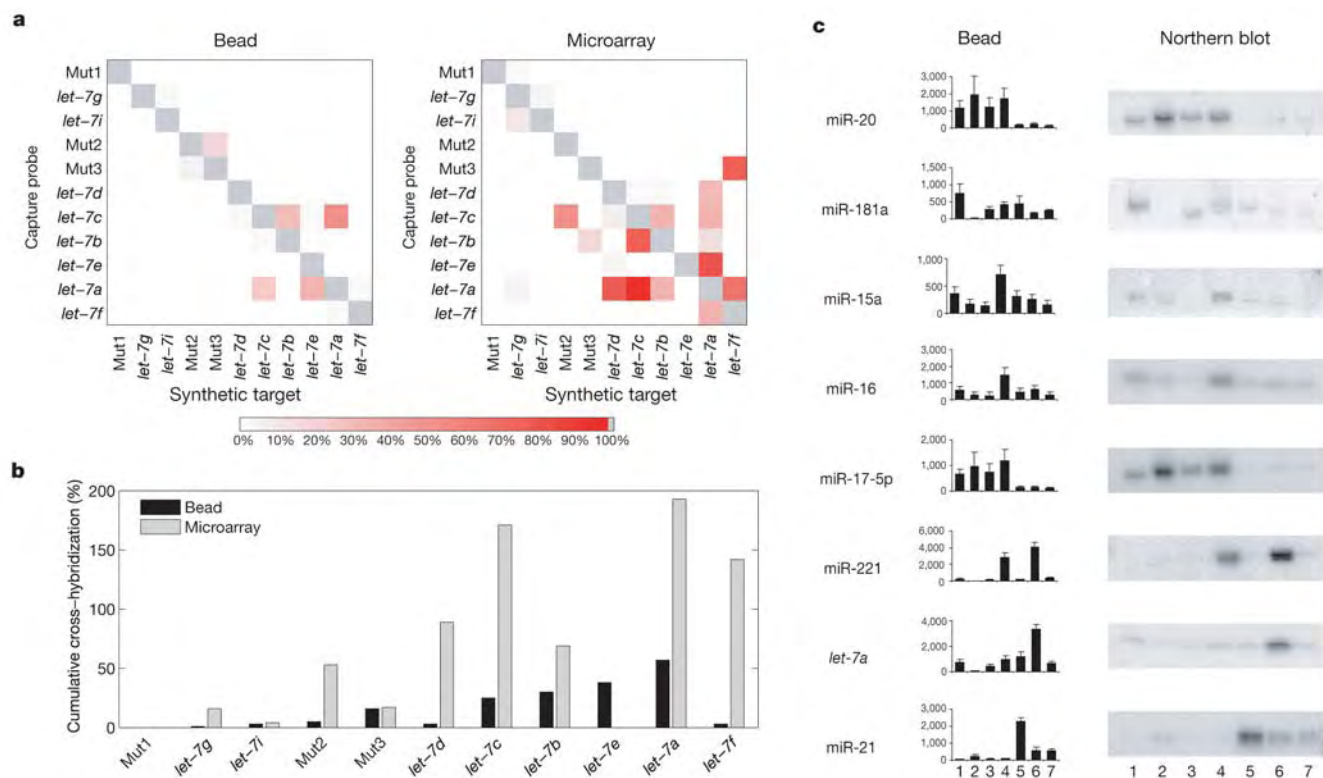


Figure 1 | Specificity and accuracy of bead-based miRNA detection.

a, Synthetic oligonucleotides corresponding to the *let-7* family and mutants (mut1–3) (see Supplementary Information for sequence similarity) were PCR-labelled and hybridized separately on beads and a glass microarray. Synthetic targets are indicated on the horizontal axis, capture probes on the vertical axis. Values represent the proportion of signal relative to correct

probe (set to 100%). **b**, Cumulative cross-hybridization on capture probes. **c**, Northern blot versus bead detection of cell line samples (lanes 1–7: HEL, K562, TF-1, 293, MCF-7, PC-3, SKMEL-5 cells). Bead results shown on the left are averages from three independent experiments for HEL, TF-1, 293, MCF-7 and PC-3 cells, and two experiments for K562 and SKMEL-5 cells. Error bars indicate s.d.

Supplementary Information). These experiments support the hypothesis that global changes in miRNA expression are associated with differentiation, the abrogation of which is a hallmark of all human cancers. These findings are also consistent with the recent observation that mouse embryonic stem cells lacking Dicer, an enzyme required for miRNA maturation, fail to differentiate normally²².

We next turned to a more challenging diagnostic distinction: that of tumours of histologically uncertain cellular origin. It is estimated that 2–4% of all cancer diagnoses represent cancers of unknown origin or diagnostic uncertainty (see ref. 23). To address this, we

analysed 17 poorly differentiated tumours, the histological appearance of which was non-diagnostic, but for which clinical diagnosis was established by anatomical context, either directly (for example, a primary tumour arising in the colon) or indirectly (a metastasis of a previously identified primary tumour). A training set of 68 more-differentiated tumours, representing 11 tumour types and for which both mRNA and miRNA profiles were available, was used to generate a classifier. This classifier was then used without modification to classify the 17 poorly differentiated test samples. As a group, poorly differentiated tumours had lower global levels of miRNA expression compared with the more-differentiated training set

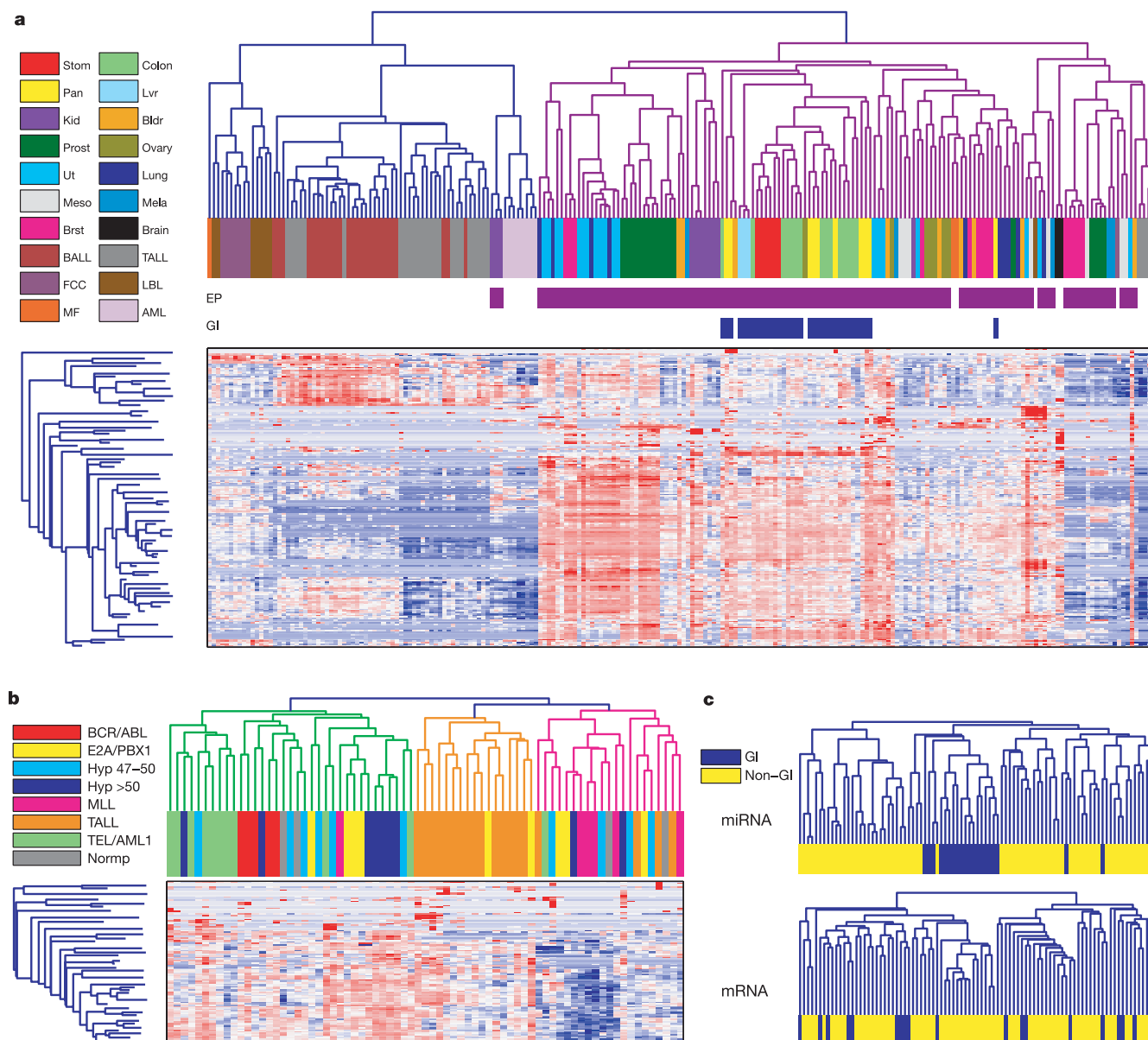


Figure 2 | Hierarchical clustering of miRNA expression. **a**, miRNA profiles of 218 samples from several different tissues were clustered (average linkage, correlation similarity). Samples are in columns, miRNAs in rows. Samples of epithelial (EP) origin or derived from the gastrointestinal tract (GI) are indicated. More detail is shown in Supplementary Fig. 4. **b**, Clustering of 73 bone marrow samples from patients with acute lymphoblastic leukaemia (ALL). Coloured bars indicate the different ALL subtypes. **c**, Comparison of miRNA data and mRNA data. For 89 epithelial samples from **a** that had mRNA expression data, hierarchical clustering was performed. Samples of GI origin are shown in blue. GI-derived samples largely cluster together in

miRNA expression space, but not in mRNA expression space. Abbreviations used: Bldr, bladder; Brst, breast; Fcc, follicular lymphoma; Kid, kidney; Lvr, liver; Mela, melanoma; Meso, mesothelioma; Pan, pancreas; Prost, prostate; Stom: stomach; Ut, uterus; AML: acute myelogenous leukaemia; BALL, B-cell ALL; LBL, diffuse large-B cell lymphoma; MF, mycosis fungoides; MLL, mixed lineage leukaemia; TALL, T-cell ALL; Hyper 47–50, hyperdiploid with 47–50 chromosomes; Hyper >50, hyperdiploid with over 50 chromosomes; Normp, normal ploidy. Further details can be found in Supplementary Information.

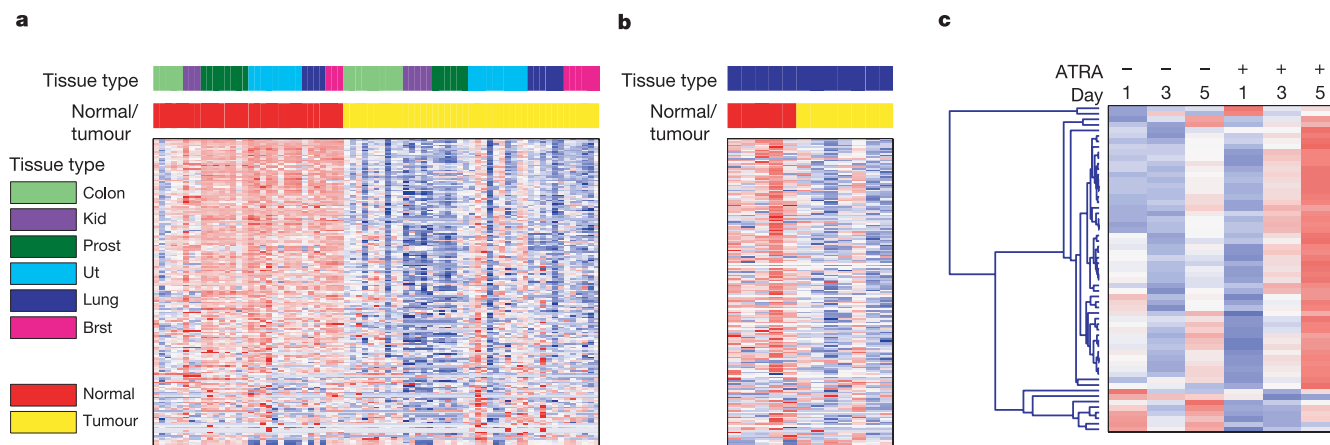


Figure 3 | Comparison between normal and tumour samples reveals global changes in miRNA expression. **a**, Markers were selected to correlate with the normal versus tumour distinction. A heatmap of miRNA expression is shown, with miRNAs sorted according to the variance-thresholded *t*-test score. **b**, miRNA markers of normal versus tumour distinction in human tissues from **a**, applied to normal lungs and lung adenocarcinomas of *K-Ras*^{LA1} mice. A *k*-nearest neighbour (*k*NN) classifier based on human

sample-derived markers yielded a perfect classification of the mouse samples (euclidean distance, $k = 3$). Mouse tumour T_MLUNG_5 (third column from right) was occasionally classified as normal using other *k*NN parameters (see Supplementary Information). **c**, HL-60 cells were treated with *all-trans* retinoic acid (ATRA +) or vehicle (–) for the indicated number of days. A heatmap of miRNA expression from a representative experiment is shown.

samples (see Supplementary Fig. 5), consistent with the notion that miRNA expression is closely linked to differentiation. Despite the overall low level of miRNA expression, the miRNA-based classifier established the correct diagnosis of the poorly differentiated samples with far greater accuracy than would be expected by chance for an 11-class classifier (12 out of 17 correct; $P < 5 \times 10^{-11}$). In contrast, the mRNA-based classifier was highly inaccurate (1 out of 17 correct; $P = 0.47$), as we previously reported⁴.

The experiments reported here demonstrate the feasibility and utility of monitoring the expression of miRNAs in human cancer. Our unexpected findings are the extraordinary level of diversity in miRNA expression across cancers, and the large amount of diagnostic information encoded in a relatively small number of miRNAs. The implication is that, unlike with mRNA expression, a modest number of miRNAs (~200 in total) might be sufficient to classify human cancers. Moreover, the bead-based miRNA detection method has the attractive property of being not only accurate and specific, but also easy to implement in a routine clinical setting. In addition, unlike mRNAs, miRNAs remain largely intact in routinely collected, formalin-fixed, paraffin-embedded clinical tissues¹⁵. More work is required to establish the clinical utility of miRNA expression in cancer diagnosis, but the work described here indicates that miRNA profiling has unexpected diagnostic potential. The mechanism by which miRNAs are under-expressed in cancer remains unknown. We did not observe substantial decreases in the mRNAs encoding components of the miRNA processing machinery (Dicer, Drosha, Argonaute2 or DGCR8 (ref. 24); see Supplementary Information), but clearly other mechanisms of miRNA regulation are possible.

The findings reported here are consistent with the hypothesis that in mammals, as in *C. elegans*, miRNAs can function to prevent cell division and drive terminal differentiation. An implication of this hypothesis is that downregulation of some miRNAs might play a causal role in the generation or maintenance of tumours. Epithelial cells affected in *C. elegans lin-4* and *let-7* miRNA mutants generate a stem-cell-like lineage, dividing to produce daughters that, like themselves, divide rather than differentiate^{5,7}. We speculate that abnormalities in miRNA expression might similarly contribute to the generation or maintenance of 'cancer stem cells', recently proposed to be responsible for cancerous growth in both leukaemias and solid tumours^{25–28}.

METHODS

Samples. Information regarding samples is available in Supplementary Table 2. Total RNAs were prepared from tissues or cell lines using TRIzol (Invitrogen), as described⁴, and in compliance with IRB protocols. Leukaemia bone marrow mononuclear cells were collected from patients treated at St. Jude Children's Research Hospital and at the Dana-Farber Cancer Institute, and their immunophenotype and genotype were determined as previously described^{29,30}. Normal mouse lung and mouse lung cancer samples were collected from *K-Ras*^{LA1} mice and genotyped as described²⁰. Lungs from four- to five-month-old mice were inflated with phosphate-buffered saline before removal. Individual lung tumours and normal lungs were dissected and immediately frozen on dry ice before RNA preparation. HL-60 cells were plated at 1.5×10^5 cells ml⁻¹ and induced to differentiate using 1 μ M *all-trans* retinoic acid (Sigma; in ethanol). Cells were harvested after 1, 3 and 5 days. Culturing conditions for other cells are detailed in the Supplementary Information.

miRNA labelling. Target preparation from total RNA followed the described procedure¹³, with modifications. Briefly, two synthetic pre-labelling-control RNA oligonucleotides (5'-pCAGUCAGUCAGUCAGUCAGUCAG-3', and 5'-pGACCUCAUGUAAACGUACAA-3', Dharmacon) were used to control for target preparation efficiency. They were each spiked at 3 fmoles per μ g total RNA. Small RNAs (18–26 nucleotides) were recovered from 1–10 μ g total RNA through denaturing polyacrylamide gel purification. Small RNAs were adaptor-ligated sequentially on the 3'-end and 5'-end using T4 RNA ligase (Amersham Biosciences). After reverse-transcription using adaptor-specific primers, products were PCR-amplified (95 °C for 40 s; 50 °C for 30 s; 72 °C for 30 s; 18 cycles for 10 μ g starting total RNA) using a 3'-primer 5'-TACTGG AATTCGCGGTTA-3' and 5' primer 5'-biotin-CAACGGAATTCCTCACTAAA-3' (IDT). For side-by-side comparison of bead detection and the glass microarray, a 5'-Alexa-532-modified primer was used for compatibility with the glass microarray. PCR products were precipitated and dissolved in 66 μ l TE buffer (10 mM TrisHCl pH 8.0, 1 mM EDTA) containing two biotinylated post-labelling-control oligonucleotides (100 fmoles of FVR506, 25 fmoles PTG20210; see Supplementary Table 1).

Bead-based detection. miRNA capture probes were 5'-amino-modified oligonucleotides with a 6-carbon linker (IDT). Capture probes for miRNAs and controls were divided into three sets (see Supplementary Table 1). To profile a sample with all probes, three assays were performed on the sample, each using one of the three probe sets. Probes were conjugated to carboxylated xMAP beads (Luminex Corporation) in 96-well plates, following the manufacturer's protocol. For each probe set, 3 μ l of every probe-bead conjugate was mixed into 1 ml of 1.5 $\times$ TMAC (4.5 M tetramethylammonium chloride, 0.15% sarkosyl, 75 mM Tris-HCl pH 8.0, 6 mM EDTA). Samples were hybridized in a 96-well plate, with two mock PCR samples (using water as template) in each plate as a background control. Hybridization was carried out overnight at 50 °C with 33 μ l of the bead

mixture and 15 μ l of labelled material. Beads were spun down, resuspended in $1 \times$ TMAC containing $10 \mu\text{g ml}^{-1}$ streptavidin-phycoerythrin (Molecular Probes) and incubated at 50°C for 10 min before data acquisition on a Luminex 100IS machine. Median fluorescence intensity values were measured.

Computational analyses. Profiling data were first scaled according to the post-labelling-controls and then the pre-labelling-controls, in order to normalize readings from different probe/bead sets for the same sample, and to normalize for the labelling efficiency, respectively, as detailed in Supplementary Methods. Data were thresholded at 32 and $\log_2$ -transformed. Hierarchical clustering was performed with average linkage and Pearson correlation. Before clustering, data were filtered to eliminate genes with expression lower than 7.25 (on a $\log_2$ scale) in all samples. Next, all features were centred and normalized to a mean of 0 and a standard deviation of 1. k -nearest-neighbour classification of normal versus tumour samples was performed with $k = 3$ in the selected feature space using euclidean distance measure. Note that different metrics were used for clustering and normal/tumour classification. Features were selected for the distinction between all normal samples versus all tumours (for colon, kidney, prostate, uterus, lung and breast; $P < 0.05$ after Bonferroni correction). P values were calculated using a variance-thresholded t -test with a minimal standard deviation of 0.75, while treating the tissue type as a confounding variable. Multi-class predictions of poorly differentiated tumours were performed using the probabilistic neural network algorithm, a gaussian-weighted nearest neighbour method. For each test sample, the tissue type that had the highest probability in multiple one-tissue-versus-the-rest predictions was assigned. Feature number and the gaussian width were optimized on the basis of leave-one-out cross-validations on the training data set. Features were selected on the basis of the variance-thresholded t -test score, requiring equal numbers of up- and down-regulated features. Distances were based on the cosine in the selected feature space.

Expression data. miRNA expression data have been submitted to the Gene Expression Omnibus (GEO; <http://www.ncbi.nlm.nih.gov/geo>) with the series accession number GSE2564. mRNA expression data were published previously⁴, and are available together with miRNA expression data at <http://www.broad.mit.edu/cancer/pub/miGCM>.

Received 2 February; accepted 5 May 2005.

- Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297 (2004).
- Ambros, V. The functions of animal microRNAs. *Nature* **431**, 350–355 (2004).
- Chung, C. H., Bernard, P. S. & Perou, C. M. Molecular portraits and the family tree of cancer. *Nature Genet.* **32**, 533–540 (2002).
- Ramaswamy, S. *et al.* Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl Acad. Sci. USA* **98**, 15149–15154 (2001).
- Ambros, V. & Horvitz, H. R. Heterochronic mutants of the nematode *Caenorhabditis elegans*. *Science* **226**, 409–416 (1984).
- Lee, R. C., Feinbaum, R. L. & Ambros, V. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* **75**, 843–854 (1993).
- Reinhart, B. J. *et al.* The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature* **403**, 901–906 (2000).
- Michael, M. Z., O'Connor, S. M., van Holst Pellekaan, N. G., Young, G. P. & James, R. J. Reduced accumulation of specific microRNAs in colorectal neoplasia. *Mol. Cancer Res.* **1**, 882–891 (2003).
- Calin, G. A. *et al.* Frequent deletions and down-regulation of micro-RNA genes *miR15* and *miR16* at 13q14 in chronic lymphocytic leukemia. *Proc. Natl Acad. Sci. USA* **99**, 15524–15529 (2002).
- Eis, P. S. *et al.* Accumulation of miR-155 and *BIC* RNA in human B cell lymphomas. *Proc. Natl Acad. Sci. USA* **102**, 3627–3632 (2005).
- Johnson, S. M. *et al.* RAS is regulated by the *let-7* microRNA family. *Cell* **120**, 635–647 (2005).
- Liu, C. G. *et al.* An oligonucleotide microchip for genome-wide microRNA profiling in human and mouse tissues. *Proc. Natl Acad. Sci. USA* **101**, 9740–9744 (2004).
- Miska, E. A. *et al.* Microarray analysis of microRNA expression in the developing mammalian brain. *Genome Biol.* **5**, R68 (2004).
- Thomson, J. M., Parker, J., Perou, C. M. & Hammond, S. M. A custom microarray platform for analysis of microRNA gene expression. *Nature Methods* **1**, 47–53 (2004).
- Nelson, P. T. *et al.* Microarray-based, high-throughput gene expression profiling of microRNAs. *Nature Methods* **1**, 155–161 (2004).
- Babak, T., Zhang, W., Morris, Q., Blencowe, B. J. & Hughes, T. R. Probing microRNAs with microarrays: tissue specificity and functional inference. *RNA* **10**, 1813–1819 (2004).
- Sun, Y. *et al.* Development of a micro-array to detect human and mouse microRNAs and characterization of expression in human organs. *Nucleic Acids Res.* **32**, e188 (2004).
- Barad, O. *et al.* MicroRNA expression detected by oligonucleotide microarrays: system establishment and expression profiling in human tissues. *Genome Res.* **14**, 2486–2494 (2004).
- Calin, G. A. *et al.* MicroRNA profiling reveals distinct signatures in B cell chronic lymphocytic leukemias. *Proc. Natl Acad. Sci. USA* **101**, 11755–11760 (2004).
- Johnson, L. *et al.* Somatic activation of the *K-ras* oncogene causes early onset lung cancer in mice. *Nature* **410**, 1111–1116 (2001).
- Stegmaier, K. *et al.* Gene expression-based high-throughput screening (GE-HTS) and application to leukemia differentiation. *Nature Genet.* **36**, 257–263 (2004).
- Kanellopoulou, C. *et al.* Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. *Genes Dev.* **19**, 489–501 (2005).
- Pavlidis, N., Briasoulis, E., Hainsworth, J. & Greco, F. A. Diagnostic and therapeutic management of cancer of an unknown primary. *Eur. J. Cancer* **39**, 1990–2005 (2003).
- Cullen, B. R. Transcription and processing of human microRNA precursors. *Mol. Cell* **16**, 861–865 (2004).
- Lapidot, T. *et al.* A cell initiating human acute myeloid leukaemia after transplantation into SCID mice. *Nature* **367**, 645–648 (1994).
- Reya, T., Morrison, S. J., Clarke, M. F. & Weissman, I. L. Stem cells, cancer, and cancer stem cells. *Nature* **414**, 105–111 (2001).
- Al-Hajj, M., Wicha, M. S., Benito-Hernandez, A., Morrison, S. J. & Clarke, M. F. Prospective identification of tumorigenic breast cancer cells. *Proc. Natl Acad. Sci. USA* **100**, 3983–3988 (2003).
- Singh, S. K. *et al.* Identification of human brain tumour initiating cells. *Nature* **432**, 396–401 (2004).
- Yeoh, E. J. *et al.* Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* **1**, 133–143 (2002).
- Ferrando, A. A. *et al.* Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia. *Cancer Cell* **1**, 75–87 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank E. Lander for critical review of the manuscript, S. Ramaswamy for discussions, and J.-P. Brunet, S. Monti, C. Ladd-Acosta and S. Shurtleff for computational help and technical assistance. We also thank J. Jacobson and Luminex Corporation for advice and technical support. E.A.M. was supported by the Howard Hughes Medical Institute. H.R.H., T.J. and T.R.G. are Investigators of the Howard Hughes Medical Institute.

Author Information miRNA expression data have been submitted to the Gene Expression Omnibus under the series accession number GSE2564. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to T.R.G. (golub@broad.mit.edu).

c-Myc-regulated microRNAs modulate E2F1 expression

Kathryn A. O'Donnell^{1,2}, Erik A. Wentzel², Karen I. Zeller³, Chi V. Dang^{1,2,3,5} & Joshua T. Mendell^{1,2,4}

MicroRNAs (miRNAs) are 21–23 nucleotide RNA molecules that regulate the stability or translational efficiency of target messenger RNAs¹. miRNAs have diverse functions, including the regulation of cellular differentiation, proliferation and apoptosis². Although strict tissue- and developmental-stage-specific expression is critical for appropriate miRNA function, mammalian transcription factors that regulate miRNAs have not yet been identified. The proto-oncogene *c-MYC* encodes a transcription factor that regulates cell proliferation, growth and apoptosis³. Dysregulated expression or function of *c-Myc* is one of the most common abnormalities in human malignancy⁴. Here we show that *c-Myc* activates expression of a cluster of six miRNAs on human chromosome 13. Chromatin immunoprecipitation experiments show that *c-Myc* binds directly to this locus. The transcription factor E2F1 is an additional target of *c-Myc* that promotes cell cycle progression^{5–7}. We find that expression of E2F1 is negatively regulated by two miRNAs in this cluster, miR-17-5p and miR-20a. These findings expand the known classes of transcripts within the *c-Myc* target gene network, and reveal a mechanism through which *c-Myc* simultaneously activates E2F1 transcription and limits its translation, allowing a tightly controlled proliferative signal.

c-Myc is a helix–loop–helix leucine zipper transcription factor that regulates an estimated 10–15% of genes in the human and *Drosophila* genomes^{7–9}. Both *c-Myc* and miRNAs have been shown to influence cell proliferation and death, and select miRNAs are known to have abnormal expression in human malignancies^{4,10}. We thus sought to determine whether *c-Myc* regulates miRNA expression.

A spotted-oligonucleotide array capable of measuring the expression of 235 human, mouse or rat miRNAs was generated and used to analyse a previously described human B-cell line, P493-6, that harbors a tetracycline-repressible *c-MYC* transgene¹¹ (Fig. 1a, b). miRNA expression profiles were analysed in tetracycline (tet)-treated (low *c-Myc*) or untreated (high *c-Myc*) cells. Six upregulated miRNAs were consistently observed in the high *c-Myc* state: miR-17-5p, miR-18, miR-19, miR-20, miR-92 and miR-106. These miRNAs are encoded by three paralogous clusters located on chromosome 13 (the *mir-17* cluster), the X chromosome (the *mir-106a* cluster) and chromosome 7 (the *mir-106b* cluster, Fig. 2a). As the array did not detect upregulation of miR-25, which is encoded by the *mir-106b* cluster, we focused our analyses on the *mir-17* and *mir-106a* clusters. Northern blotting confirmed that the miRNAs contained within these clusters were upregulated in the high *c-Myc* state (Fig. 2b). However, miR-17-3p, which has been reported to be

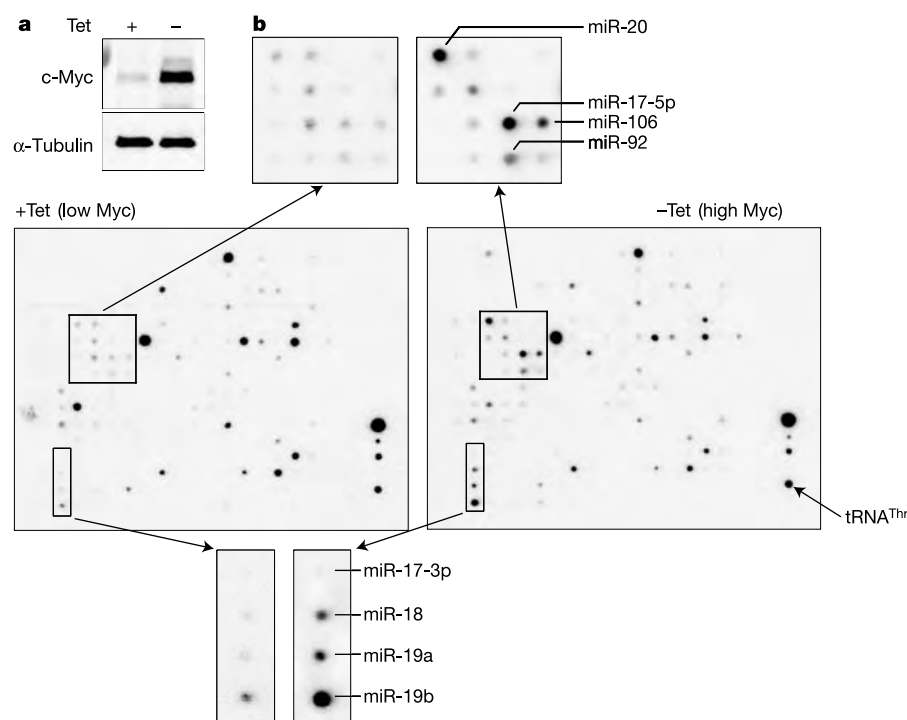


Figure 1 | MicroRNA expression profiling of P493-6 cells with high and low *c-Myc* expression.

a, Western blot analysis of *c-Myc* in untreated cells or in cells treated with tetracycline (Tet). Blots were stripped and reprobed for α -tubulin to demonstrate equal loading of samples. **b**, MicroRNA expression arrays hybridized with RNA from tet-treated (+Tet) or untreated (–Tet) cells. Magnified panels show miRNAs that were consistently upregulated in the high *c-Myc* state. A probe complementary to threonine tRNA ($tRNA^{Thr}$) served as a control for equal hybridization.

¹Program in Human Genetics and Molecular Biology, ²The Institute of Genetic Medicine, ³Departments of Medicine and ⁴Pediatrics, and ⁵the Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA.

expressed from the *mir-17* cluster, was not detectable in P493-6 cells, suggesting that it might be a miRNA* sequence (the reverse-complement strand of a miRNA; Fig. 1b and data not shown).

miRNAs are transcribed by RNA polymerase II as long primary transcripts (pri-miRNAs) that undergo sequential processing to produce mature miRNAs^{12–16}. Probes for northern blotting were designed to detect pri-miRNA transcripts from the *mir-17* and *mir-106a* clusters. These probes were complementary to unique sequence immediately upstream of the first pre-miRNA hairpin in each cluster. The *mir-17* cluster-specific probe detected three transcripts (approximately 3.2, 1.3 and 0.8 kilobases (kb) in size) that were induced in the high c-Myc state (Fig. 2c). It has been previously reported that the *mir-17* cluster is contained within an alternatively spliced host transcript termed *C13orf25* (ref. 17). The observed transcripts represent alternatively spliced 5'-cleavage products of *C13orf25* that remain following excision of pre-miRNAs (our unpublished observations). A similar probe complementary to sequence immediately upstream of the *mir-106a* cluster did not detect any transcripts in P493-6 cells (data not shown). These data demonstrate that the *mir-17* cluster is upregulated in the high c-Myc state.

In order to confirm that regulation of the *mir-17* cluster by c-Myc was not restricted to P493-6 cells, we examined levels of miR-18 and miR-20 in previously described wild-type rat fibroblasts (TGR), rat fibroblasts containing a homozygous deletion of *c-Myc* (HO15.19), or *c-Myc* null fibroblasts reconstituted with wild-type c-Myc (HO15.19-MYC)^{18,19}. miR-18 and miR-20 were expressed at approximately 50% of wild-type levels in the absence of c-Myc, but wild-type expression levels of these miRNAs were restored in the c-Myc-reconstituted null cells (Fig. 2d).

We next performed chromatin immunoprecipitation (ChIP) experiments in P493-6 cells to determine whether human c-Myc binds directly to the *mir-17* cluster genomic locus. First, 10 kb of sequence on chromosome 13 surrounding the *mir-17* cluster was examined for putative c-Myc-binding sites. c-Myc is known to bind to the canonical E-box sequence CACGTG, as well as to non-canonical sequences including CATGTG²⁰. We identified seven putative binding sites matching these sequences. Four of these sites were conserved between human and mouse, and located within a 30-base-pair window of at least 65% nucleotide identity between these species (Fig. 3a, labelled in red). Real-time polymerase chain reaction (PCR) amplicons were designed to assay for all putative binding sites (both conserved and non-conserved) in ChIP samples. Background signals were very low at all tested amplicons in negative control samples immunoprecipitated without antibody or with an antibody directed against hepatocyte growth factor (HGF), which is not expressed in these cells (data not shown). We obtained clear evidence for *in vivo* association of c-Myc with a region containing a conserved CATGTG sequence 1,480 nucleotides upstream of *mir-17-5p* (Fig. 3b, amplicon 3). This site is located in intron 1 of *C13orf25*. c-Myc is known to frequently bind to sites in the first intron of its transcriptional target genes²⁰. As we were not able to design amplicons between nucleotides –1500 and –3280 (owing to the presence of a CpG island that prevented efficient amplification), we cannot rule out the possibility that c-Myc also binds within this region. Our data demonstrate that c-Myc binds directly to the *mir-17* cluster locus, providing strong evidence that these miRNAs are directly regulated by this transcription factor.

Seven putative binding sites in the vicinity of the *mir-106a* cluster were also identified and assayed for c-Myc binding. No ChIP signals were observed, consistent with the northern blot data showing an absence of detectable transcripts produced from this locus in P493-6 cells (data not shown).

The behaviour of the *mir-17* cluster was also examined during serum stimulation in primary human fibroblasts. Serum deprivation followed by serum stimulation of fibroblasts results in a transient induction of c-Myc²¹ (Fig. 3c). Real-time PCR analysis demonstrates

that under these conditions, expression of the *mir-17* host transcript is induced with similar kinetics (Fig. 3d). Consistent with the behaviour of other known c-Myc target genes, expression levels remain elevated after c-Myc levels decrease²². Furthermore, ChIP analysis demonstrates that association of c-Myc with the *mir-17* genomic locus mirrors c-Myc expression and coincides with induction of miRNA cluster expression (Fig. 3e). These results provide further evidence that the *mir-17* cluster is directly regulated by c-Myc, and show that induction of these miRNAs is a physiologic response to growth stimuli.

To study the functional consequences of induction of the *mir-17* cluster by c-Myc, we examined mRNAs that are predicted targets of these miRNAs. The transcription factor E2F1, which is predicted to be regulated by miR-17-5p and miR-20a (ref. 23), was initially chosen for further analysis. E2F1 expression promotes G1-to-S phase progression in mammalian cells by activating genes involved in DNA replication and cell cycle control⁵. Expression of the *E2F1* gene is known to be induced by c-Myc^{6,7}. c-Myc expression is also induced by E2F1, revealing a putative positive feedback circuit²¹. We hypothesized that negative regulation of *E2F1* translation by miR-17-5p and miR-20a provides a mechanism to dampen this reciprocal activation, promoting tightly controlled expression of c-Myc and E2F1 gene products.

To determine whether *E2F1* is a target of miR-17-5p and miR-20a,

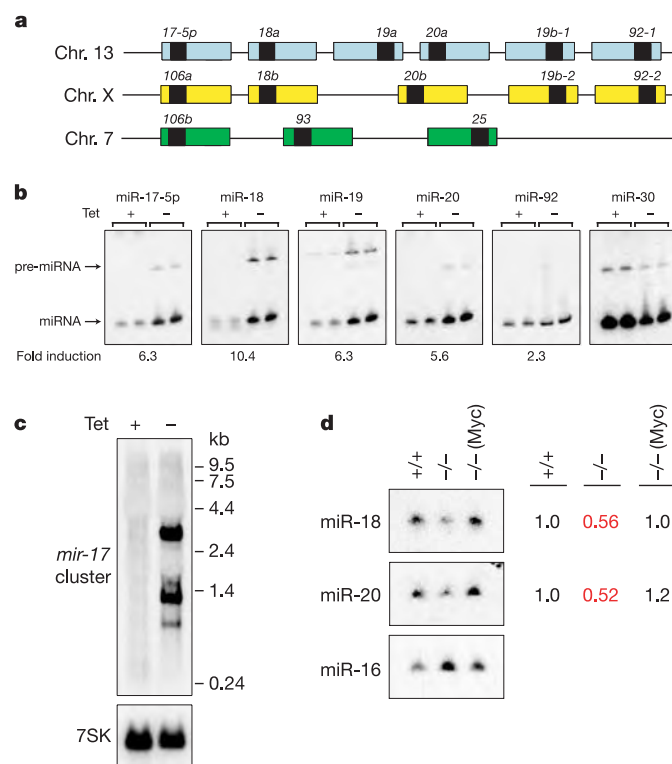


Figure 2 | c-Myc induces expression of the miR-17 cluster. **a**, Schematic representation of the *mir-17*, *mir-106a* and *mir-106b* clusters. *mir-18b* and *mir-20b* are predicted on the basis of homology to *mir-18a* and *mir-20a*, respectively³⁰. **b**, Northern blot analysis of miRNAs in P493-6 cells. Duplicate samples are shown, and miR-30 served as a loading control. Blots were also probed for miR-16 and miR-29 as loading controls, and similar results were obtained (data not shown). **c**, Northern blot analysis of total RNA from P493-6 cells with a probe specific for the *mir-17* cluster. 7SK RNA served as a loading control. **d**, Northern blot analysis of miRNAs in wild-type rat fibroblasts (+/+), rat fibroblasts with a homozygous deletion of *c-Myc* (–/–), or knockout fibroblasts reconstituted with wild-type c-Myc (–/– (Myc)). Quantification of radioactive signal intensity is shown on the right.

we used HeLa cells, which naturally express the *mir-17* cluster. 2'-O-methyl antisense oligoribonucleotides, which have been shown to block miRNA function^{24,25}, were designed to inhibit miR-17-5p and miR-20a. To monitor the degree of miRNA inhibition, we generated sensor constructs with sites perfectly complementary to miR-17-5p or miR-20a in the 3'-untranslated region (UTR) of firefly luciferase. When introduced into HeLa cells, these constructs showed an 80–90% reduction in luciferase activity compared with control constructs containing the reverse-complement sequence of the miRNA-binding sites; this demonstrates efficient downregulation by endogenous miR-17-5p and miR-20a. Co-transfection of these plasmids with miR-17-5p or miR-20a antisense oligonucleotides, but not scrambled oligonucleotides, enhanced expression of the sensor constructs, indicating inhibition of these miRNAs (Fig. 4a). Because of nucleotide similarity between miR-17-5p and miR-20a, both were inhibited by antisense oligonucleotides directed against either miRNA. Transfection with miR-17-5p and miR-20a antisense oligonucleotides, but not scrambled oligonucleotides, resulted in an approximately fourfold increase in E2F1 protein levels without altering *E2F1* mRNA abundance (Fig. 4b, c).

We also determined the consequence of overexpressing the *mir-17* cluster on E2F1 expression. The entire *mir-17* cluster and approximately 150 nucleotides of flanking sequence were cloned into a mammalian expression vector, under the control of a cytomegalovirus (CMV) promoter. When transfected into HeLa cells, this construct (CMV-*mir-17* cluster) produces the appropriately processed miRNAs, as assessed by northern blotting (Fig. 4d and not shown). Transient overexpression of these miRNAs resulted in a 50% decrease in E2F1 protein levels (Fig. 4e) without affecting *E2F1* mRNA abundance (Fig. 4f).

To demonstrate that miR-17-5p and miR-20a directly regulate E2F1 expression, luciferase reporter constructs containing a portion of the *E2F1* 3'-UTR were generated and mutations were introduced into the predicted miRNA-binding sites (see Supplementary

Fig. 1a, b). The mutant construct yielded approximately threefold higher luciferase expression compared with the wild-type construct when transfected into HeLa cells, providing evidence that the endogenously expressed miRNAs decrease E2F1 expression by recognizing these sites (see Supplementary Fig. 1c).

Last, we examined *E2F1* mRNA and protein levels in P493-6 cells with high and low c-Myc expression (leading to high and low expression of the *mir-17* cluster, respectively). Consistent with previously reported data^{6,7}, c-Myc potently induces *E2F1* mRNA (Fig. 4g). Remarkably, E2F1 protein levels were only modestly induced under these conditions, suggesting a greatly reduced translational yield (Fig. 4h). Taken together with the results from HeLa cells, these data support a model in which miR-17-5p and miR-20a limit c-Myc-mediated induction of E2F1 expression, preventing uncontrolled reciprocal activation of these gene products. As E2F1 protein is known to accumulate late in G1, and c-Myc (and consequently the *mir-17* cluster) are activated early in G1 (refs 5, 21), we speculate that E2F1 translational efficiency is decreased, but not completely inhibited, by these miRNAs during normal cell-cycle progression. Consistent with a dampened translational efficiency, E2F1 protein accumulation is delayed relative to *E2F1* mRNA induction during a time course of serum stimulation in primary fibroblasts. In contrast, c-Myc protein levels closely mirror mRNA levels under these conditions (see Supplementary Fig. 2). As several other documented c-Myc target genes are also predicted targets of the *mir-17* cluster (for example, *RPS6KA5*, *BCL11B*, *PTEN* and *HCFC2*)^{20,23}, a widespread mechanism may exist through which c-Myc and other transcription factors precisely control expression of target genes by simultaneously activating their transcription and limiting their translation.

These results identify miRNAs as targets of c-Myc, expanding the known classes of transcripts within the c-Myc target gene network. Furthermore, they suggest that the *mir-17* cluster, by decreasing E2F1 expression, tightly regulates c-Myc-mediated cellular proliferation.

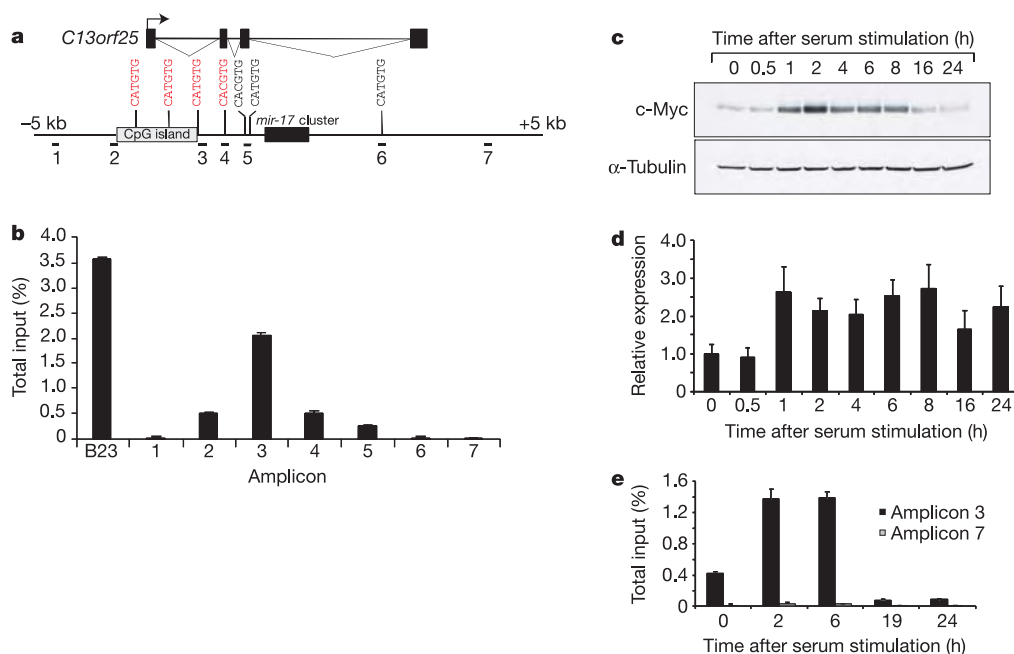


Figure 3 | c-Myc binds directly to the *mir-17* cluster genomic locus. **a**, Schematic representation of the genomic interval encompassing the *mir-17* cluster. Putative c-Myc binding sites are indicated (CACGTG or CATGTG); those in red are conserved between human and mouse. The location and structure of *C13orf25* is indicated. Real-time PCR amplicons are represented by numbered lines. **b**, Real-time PCR analysis of c-Myc chromatin immunoprecipitates. Amplification of a validated c-Myc-binding

site in intron 1 of the B23 gene served as a positive control²². **c**, Western blot analysis of c-Myc protein levels following serum stimulation of primary human fibroblasts. **d**, Real-time PCR analysis of *mir-17* cluster expression (**d**) and c-Myc chromatin immunoprecipitates in serum-stimulated fibroblasts (**e**). Error bars for all panels represent standard deviations derived from at least three independent measurements.

In this context, these miRNAs might have tumour suppressor activity. Accordingly, loss-of-heterozygosity of the chromosomal region encompassing the *mir-17* cluster (13q31) has been observed in human malignancies²⁶. However, amplification of this region and

overexpression of *C13orf25*, the host transcript of the *mir-17* cluster, has been described in diffuse large B-cell lymphoma¹⁷, and miR-19a and miR-92-1 have been shown to be upregulated in B-cell chronic lymphocytic leukaemia²⁷. These observations suggest that miRNAs might also possess oncogenic activity. It is thus likely that these miRNAs influence cell proliferation and tumorigenesis in a cell-type-specific manner, depending on the milieu of target mRNAs that are expressed. The results described here provide an experimental framework for further functional dissection of this miRNA cluster, in order to fully delineate its role in normal cellular physiology and malignancy.

METHODS

Tissue culture. P493-6 cells (from D. Eick) were cultured in RPMI 1640 media supplemented with 10% fetal bovine serum (FBS), penicillin and streptomycin. To repress *c-MYC* expression, cells were grown in the presence of $0.1 \mu\text{g ml}^{-1}$ tetracycline (Sigma) for 72 h. TGR-1 cells (wild-type rat fibroblasts) and HO15.19 cells (*c-Myc*^{-/-} rat fibroblasts) were a gift from J. Sedivy. Rat cells, HeLa cells, and primary human fibroblasts (obtained from ATCC) were grown in Dulbecco's Modified Eagle's Medium (DMEM) with 10% FBS, penicillin and streptomycin. For serum stimulation experiments, primary fibroblasts were grown in DMEM with 0.1% FBS for 48 h. DMEM containing 10% FBS was then added and cells were collected at the indicated time points.

miRNA expression profiling. miRNA expression arrays were generated and probed essentially as described²⁸, with the following modifications. Oligonucleotide probes (each a concatenated triple repeat of sequence antisense to a mature miRNA) were synthesized and spotted at a concentration of $10 \mu\text{M}$ on GeneScreen Plus membranes (Perkin Elmer) with a 96-pin spotter (V&P Scientific). In addition to 235 miRNA probes, 25 probes each containing three mismatches were spotted to assess hybridization specificity. These probes produced significantly less signal intensity (in general, an approximately tenfold decrease) compared with wild-type probes. A probe complementary to tRNA^{Thr} was used to control for hybridization efficiency. All probe sequences are provided in Supplementary Table 1. Total RNA ($10 \mu\text{g}$) was isolated with Trizol Reagent (Invitrogen) and samples were enriched for low molecular weight RNA using a Microcon YM-100 column (Amicon). The resulting RNA was end-labelled with γ -[³²P]-ATP using the KinaseMax kit (Ambion). After removing unincorporated radionucleotide using Microspin G-25 columns (Amersham), labelled RNA was hybridized to membranes using MicroHyb buffer (Invitrogen). Signals were quantified using a Personal FX phosphorimager (Bio-Rad).

Northern blot analysis. For miRNA northern blots, $20 \mu\text{g}$ of total RNA was separated on 15% denaturing polyacrylamide gels, electrotransferred to GeneScreen Plus membranes, and hybridized using UltraHyb-Oligo buffer (Ambion). Oligonucleotides complementary to mature miRNAs, and end-labelled with T4 Kinase (Invitrogen), were used as probes. Probe sequences were as follows: miR-17-5p, 5'-ACTACCTGCAGTGAAGCACTTTG-3'; miR-18a, 5'-TATCTGCACTAGATGCACCTTA-3'; miR-19a, 5'-TCAGTTTTCATAGATTTCACACA-3'; miR-20a, 5'-CTACCTGCAGTATAAGCACTTTA-3'; miR-92, 5'-ACAGGCCGGACAAGTGAATA-3'; miR-30c, 5'-GCTGAGAGTGAGGATGTTTACA-3'; miR-16, 5'-CGCCAATATTTACGTGCTGCTA-3'.

For conventional northern blotting, $20 \mu\text{g}$ of total RNA was separated on 1.2% formaldehyde-agarose gels, transferred to GeneScreen Plus, and hybridized with randomly primed labelled probes using UltraHyb buffer. Probes were generated by PCR using the following primers: *mir-17* cluster probe, sense 5'-ACATGGAC TAAATTGCCCTTTAAATG-3', antisense 5'-AATCTTCAGTTTACAAGGTG ATG-3'; *mir-106a* cluster probe, sense 5'-CATCTGGGTTTACATGCTCC-3', antisense 5'-CAAAATTTTAAGTCTTCCAGGAGC-3'; 7SK RNA probe, sense 5'-GACATCTGTACCCCATTTGATC-3', antisense 5'-TCTGCAGTC TTGGAAGCTTGAC-3'; *E2F1* probe, sense 5'-TGTGTGATGATGCCAT GTGTG-3', antisense 5'-GCAAATCAAAGTGCAGATTGGAG-3'.

Western blot analysis. Antibodies for immunoblotting were as follows: anti-*c-Myc* mouse monoclonal (clone 9E10; Zymed), anti-E2F1 mouse monoclonal (clones KH20 and KH95; Upstate) and anti- α -tubulin mouse monoclonal (Calbiochem). Scanned images were quantified using Quantity One software (Bio-Rad).

Chromatin immunoprecipitation and real-time PCR. Cells were crosslinked with formaldehyde, and chromatin was immunoprecipitated as previously described²⁹. Rabbit polyclonal *c-Myc* (sc-764, Santa Cruz Biotechnology) and human HGF antibodies (sc-7949, Santa Cruz) were used to immunoprecipitate chromatin fragments. Real-time PCR was performed using an ABI 7700 Sequence Detection System with the SYBR Green PCR core reagent kit (Perkin Elmer Applied Biosystems). Sequences of primers used to amplify ChIP samples

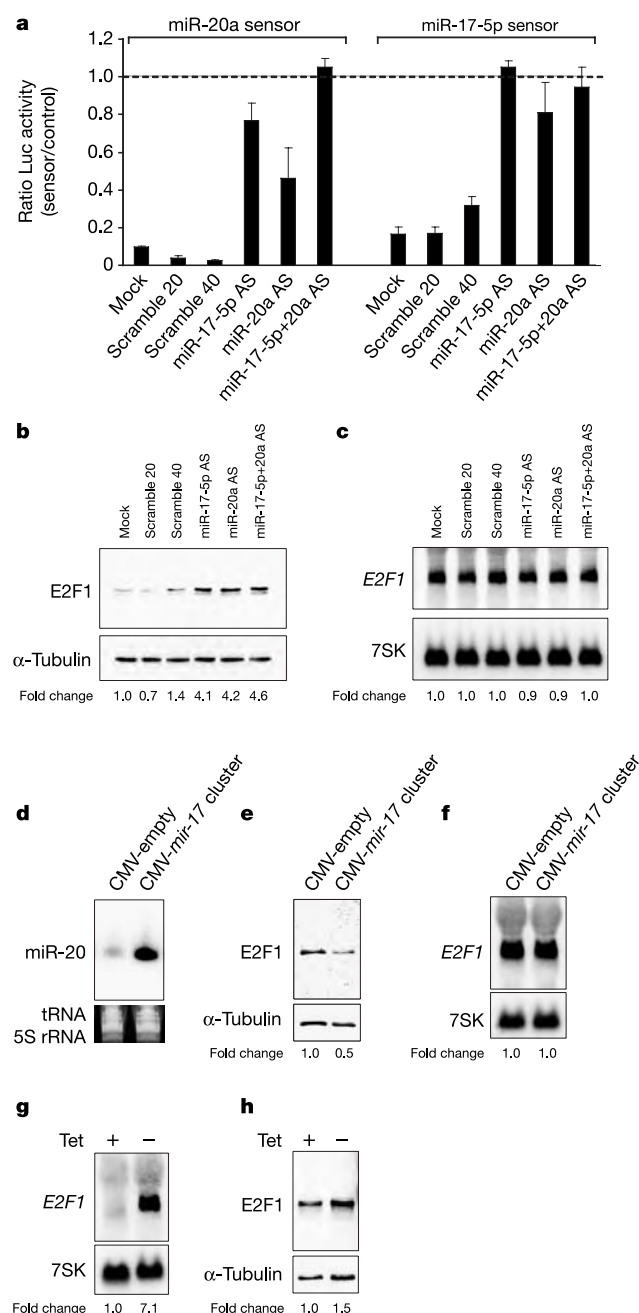


Figure 4 | miR-17-5p and miR-20a regulate E2F1 translational yield.

a, Inhibition of miR-17-5p and miR-20a by 2'-O-methyl-oligoribonucleotides. Sensor or control luciferase constructs were transfected into HeLa cells alone (mock) or with the following oligonucleotides: scrambled nucleotide at 20 or 40 pmol, or 20 pmol of miR-17-5p or miR-20a antisense (AS), either individually (miR-17-5p AS or miR-20a AS) or pooled (miR-17-5p + 20a AS). The ratio of normalized sensor to control luciferase activity is shown. Error bars represent standard deviations. **b**, **c**, Western blot (**b**) and northern blot (**c**) analysis of E2F1 in antisense-treated HeLa cells. **d**, Northern blot analysis of miR-20 in transfected HeLa cells. **e**, **f**, Western blot (**e**) and northern blot (**f**) analysis of E2F1 in transfected HeLa cells. **g**, **h**, Northern blot (**g**) and western blot (**h**) analysis of E2F1 in P493-6 cells. Fold changes shown are mean values derived from three experiments.

are provided in Supplementary Table 2. For quantification of the *C13orf25* transcript by real-time PCR, primers for amplicon 5 (Supplementary Table 2) were used. Reactions lacking reverse transcriptase were performed in parallel, to ensure that amplified fragments were derived from complementary DNA.

Oligoribonucleotides, sensor plasmids and luciferase assays. 2'-O-methyl oligoribonucleotides (scramble, 5'-AAAACCUUUUGACCGAGCGUGUU-3'; miR-17-5p antisense, 5'-ACUACCUGCACUGUAAGCACUUUG-3'; miR-20a antisense, 5'-CUACCUGCACUAUAAGCACUUUA-3') were synthesized by Integrated DNA Technologies. Sensor and control luciferase constructs were made by ligating oligonucleotides containing two sites with perfect complementarity to miR-17-5p or miR-20a into the *Xba*I site of the pGL3-control vector (Promega). Twenty-four hours before transfection, HeLa cells were plated at 150,000 cells per well in a 24-well plate. Sensor or control plasmid (200 ng) plus 80 ng pRL-SV40 (Promega) were transfected alone or in combination with 20 or 40 pmol of 2'-O-methyl oligoribonucleotides, using Lipofectamine 2000 (Invitrogen). Luciferase assays were performed 24 h after transfection using the Dual Luciferase Reporter Assay System (Promega). Firefly luciferase activity was normalized to *renilla* luciferase activity for each transfected well. Two independent plasmid preparations were each transfected at least three times (on different days). Each transfected well was assayed in triplicate.

For analysis of E2F1 mRNA and protein levels, 200 pmol 2'-O-methyl oligoribonucleotides were transfected into HeLa cells growing in 6-well dishes (plated at 170,000 cells per well 24 h before transfection) using oligofectamine (Invitrogen). RNA and protein were collected 72 h after transfection.

Overexpression of the *mir-17* cluster. The *mir-17* cluster was amplified from genomic DNA and cloned into pcDNA3.1/V5-His-TOPO (Invitrogen). The following primers were used: sense 5'-CTAATGGACCTCATATCTTTGAG-3'; antisense 5'-GAAACAAGACAAGATGTATTACAC-3'. The correct sequence of the amplified product was confirmed by sequencing. The expression plasmid was transfected into HeLa cells using HeLa Monster (Mirus).

Received 14 December 2004; accepted 20 April 2005.

- Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297 (2004).
- Ambros, V. The functions of animal microRNAs. *Nature* **431**, 350–355 (2004).
- Levens, D. Disentangling the MYC web. *Proc. Natl Acad. Sci. USA* **99**, 5757–5759 (2002).
- Cole, M. D. & McMahon, S. B. The Myc oncoprotein: a critical evaluation of transactivation and target gene regulation. *Oncogene* **18**, 2916–2924 (1999).
- Bracken, A. P., Ciro, M., Cocito, A. & Helin, K. E2F target genes: unraveling the biology. *Trends Biochem. Sci.* **29**, 409–417 (2004).
- Leone, G., DeGregori, J., Sears, R., Jakoi, L. & Nevins, J. R. Myc and Ras collaborate in inducing accumulation of active cyclin E/Cdk2 and E2F. *Nature* **387**, 422–426 (1997).
- Fernandez, P. C. *et al.* Genomic targets of the human c-Myc protein. *Genes Dev.* **17**, 1115–1129 (2003).
- Li, Z. *et al.* A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. *Proc. Natl Acad. Sci. USA* **100**, 8164–8169 (2003).
- Orian, A. *et al.* Genomic binding by the *Drosophila* Myc, Max, Mad/Mnt transcription factor network. *Genes Dev.* **17**, 1101–1114 (2003).
- McManus, M. T. MicroRNAs and cancer. *Semin. Cancer Biol.* **13**, 253–258 (2003).
- Pajic, A. *et al.* Cell cycle activation by c-myc in a Burkitt lymphoma model cell line. *Int. J. Cancer* **87**, 787–793 (2000).
- Lee, Y. *et al.* MicroRNA genes are transcribed by RNA polymerase II. *EMBO J.* **23**, 4051–4060 (2004).
- Cai, X., Hagedorn, C. H. & Cullen, B. R. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA* **10**, 1957–1966 (2004).
- Rodriguez, A., Griffiths-Jones, S., Ashurst, J. L. & Bradley, A. Identification of mammalian microRNA host genes and transcription units. *Genome Res.* **14**, 1902–1910 (2004).
- Lee, Y. *et al.* The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425**, 415–419 (2003).
- Cullen, B. R. Transcription and processing of human microRNA precursors. *Mol. Cell* **16**, 861–865 (2004).
- Ota, A. *et al.* Identification and characterization of a novel gene, *C13orf25*, as a target for 13q31-q32 amplification in malignant lymphoma. *Cancer Res.* **64**, 3087–3095 (2004).
- Guo, Q. M. *et al.* Identification of c-myc responsive genes using rat cDNA microarray. *Cancer Res.* **60**, 5922–5928 (2000).
- Mateyak, M. K., Obaya, A. J., Adachi, S. & Sedivy, J. M. Phenotypes of c-Myc-deficient rat fibroblasts isolated by targeted homologous recombination. *Cell Growth Differ.* **8**, 1039–1048 (1997).
- Zeller, K. I., Jegga, A. G., Aronow, B. J., O'Donnell, K. A. & Dang, C. V. An integrated database of genes responsive to the Myc oncogenic transcription factor: identification of direct genomic targets. *Genome Biol.* **4**, R69 (2003).
- Matsumura, I., Tanaka, H. & Kanakura, Y. E2F1 and c-Myc in cell growth and death. *Cell Cycle* **2**, 333–338 (2003).
- Zeller, K. I. *et al.* Characterization of nucleophosmin (B23) as a Myc target by scanning chromatin immunoprecipitation. *J. Biol. Chem.* **276**, 48285–48291 (2001).
- Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P. & Burge, C. B. Prediction of mammalian microRNA targets. *Cell* **115**, 787–798 (2003).
- Meister, G., Landthaler, M., Dorsett, Y. & Tuschl, T. Sequence-specific inhibition of microRNA- and siRNA-induced RNA silencing. *RNA* **10**, 544–550 (2004).
- Hutvagner, G., Simard, M. J., Mello, C. C. & Zamore, P. D. Sequence-specific inhibition of small RNA function. *PLoS Biol.* **2**, E98 (2004).
- Lin, Y. W. *et al.* Loss of heterozygosity at chromosome 13q in hepatocellular carcinoma: identification of three independent regions. *Eur. J. Cancer* **35**, 1730–1734 (1999).
- Calin, G. A. *et al.* MicroRNA profiling reveals distinct signatures in B cell chronic lymphocytic leukemias. *Proc. Natl Acad. Sci. USA* **101**, 11755–11760 (2004).
- Krichevsky, A. M., King, K. S., Donahue, C. P., Khrapko, K. & Kosik, K. S. A microRNA array reveals extensive regulation of microRNAs during brain development. *RNA* **9**, 1274–1281 (2003).
- Boyd, K. E., Wells, J., Gutman, J., Bartley, S. M. & Farnham, P. J. c-Myc target gene specificity is determined by a post-DNA binding mechanism. *Proc. Natl Acad. Sci. USA* **95**, 13887–13892 (1998).
- Tanzer, A. & Stadler, P. F. Molecular evolution of a microRNA cluster. *J. Mol. Biol.* **339**, 327–335 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors thank D. Eick and J. Sedivy for cell lines, J. Kim for assistance with ChIP, and F. Spencer and M. Awad for critical reading of the manuscript. This work was supported in part by a grant from the NIH (C.V.D.) and a Sidney Kimmel Pilot Project Grant (J.T.M.). C.V.D. is the Johns Hopkins Family Professor in Oncology Research, and J.T.M. is a March of Dimes Basil O'Connor Scholar.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.T.M. (jmendell@jhmi.edu).

LETTERS

Correlation of structural elements and infectivity of the HET-s prion

Christiane Ritter^{1*}, Marie-Lise Maddelein^{2*}, Ansgar B. Siemer^{3*}, Thorsten Lührs¹, Matthias Ernst³, Beat H. Meier³, Sven J. Saupe² & Roland Riek¹

Prions are believed to be infectious, self-propagating polymers of otherwise soluble, host-encoded proteins^{1,2}. This concept is now strongly supported by the recent findings that amyloid fibrils of recombinant prion proteins from yeast^{3–5}, *Podospora anserina*⁶ and mammals⁷ can induce prion phenotypes in the corresponding hosts. However, the structural basis of prion infectivity remains largely elusive because acquisition of atomic resolution structural properties of amyloid fibrils represents a largely unsolved technical challenge. HET-s, the prion protein of *P. anserina*, contains a carboxy-terminal prion domain comprising residues 218–289. Amyloid fibrils of HET-s(218–289) are necessary and sufficient for the induction and propagation of prion infectivity⁶. Here, we have used fluorescence studies, quenched hydrogen exchange NMR and solid-state NMR to determine the sequence-specific positions of amyloid fibril secondary structure elements of HET-s(218–289). This approach revealed four β -strands constituted by two pseudo-repeat sequences, each forming a β -strand-turn- β -strand motif. By using a structure-based mutagenesis approach, we show that this conformation is the functional and infectious entity of the HET-s prion. These results correlate distinct structural elements with prion infectivity.

The prion form of the protein HET-s is involved in a programmed cell death phenomenon termed heterokaryon incompatibility^{8,9}. This reaction occurs in filamentous fungi when cells of incompatible genotype fuse and form a mixed cell. In *P. anserina*, two incompatible genotypes, called *het-s* and *het-S*, encode the proteins HET-s and HET-S, respectively. They are both 289 amino acids long and differ in only 13 residues¹⁰. However, only HET-s can form a prion¹¹: *P. anserina* cells expressing the HET-s protein exist either in a prion state called [Het-s], or in a non-prion state called [Het-s*]. Infection of a [Het-s*] cell by a [Het-s] cell occurs by simple contact. The cell death reaction is triggered if a cell in the [Het-s] prion state fuses with a cell expressing HET-S. In contrast, a prion-free [Het-s*] cell can readily fuse with a cell expressing HET-S to form a viable mixed cell. Hence, the formation of the [Het-s] prion can easily be assayed by measuring its function in heterokaryon incompatibility. Both HET-s and the prion domain fragment HET-s(218–289) form amyloid fibrils *in vitro* that are infectious to [Het-s*] cells of *P. anserina*^{6,12,13}. Transition to the prion state *in vivo* can also be detected by following the aggregation of a HET-s–green fluorescent protein (GFP) fusion protein^{14,15}.

To elucidate the fold of the amyloid fibrils of HET-s(218–289), we first identified the sequence-specific positions of regular secondary structural elements using an improved technique of quenched hydrogen exchange measured by solution NMR¹⁶ (see Methods). This technique allows the identification of solvent-protected

backbone amide protons involved in hydrogen bonds. The [¹⁵N,¹H] correlation NMR spectrum (Fig. 1a) contained one assigned cross-peak for each backbone amide of HET-s(218–289), with the exception of 289, enabling a residue-specific determination of the hydrogen exchange rates. Upon exchange in D₂O buffer for 6 weeks, the intensity of about 45% of the resonances was significantly reduced or absent from the spectrum (Fig. 1b). This suggested that the corresponding amides have undergone exchange with solvent deuterons, which are not visible in this experiment.

The hydrogen exchange was followed over a total period of 3 months. All residues displayed a monoexponential decay (Supplementary Fig. S1), indicating that the structure of the fibrils was well defined and homogeneous. The summarized hydrogen exchange data (Fig. 1e) show that owing to exchange rates faster than 5 h^{–1}, the eight amino-terminal residues, the five carboxy-terminal residues and residues 247–261 are only weakly or not protected and may therefore be conformationally disordered. Four segments were observed that displayed slow exchange rates of 10^{–2} h^{–1} to 10^{–5} h^{–1} and were thus considered to be involved in hydrogen bonds. They comprise residues 226–234, 236–246, 262–270 and 272–282. Similarly slow exchange rates have been observed for A β (1–42) fibrils¹⁷ and reflect the extraordinary stability and compactness amyloid fibrils can achieve. Because circular dichroism experiments have established that the HET-s(218–289) fibrils contain mainly β -sheet secondary structure¹³, it is likely that the four protected segments represent four distinct β -strands. Hydrogen exchange data were also accumulated for amyloid fibrils of the full-length HET-s protein. The protection pattern of the peptide segment 218–288 in HET-s was essentially identical to that of HET-s(218–289) (Supplementary Fig. S2). This suggests that the prion domain acts as an independent folding unit.

To determine which residues are actually involved in β -sheet secondary structure, high-resolution solid-state NMR spectra of the HET-s(218–289) fibrils were recorded. The ¹³C–¹³C homonuclear DREAM¹⁸ correlation spectrum of uniformly ¹⁵N and ¹³C isotope-labelled amyloid fibrils is shown in Fig. 1c, d. The resonance lines are strikingly narrow (down to 0.2 p.p.m.) and are comparable to those of micro-crystalline proteins, indicating a highly ordered atomic structure for part of the fibrils. Sequence-specific chemical shift assignment could be obtained for residues 226–248 and 262–282 (with the exception of 274). The missing resonances of the remaining residues are probably a consequence of dynamical or structural heterogeneity¹⁹. Deviations of the combined ¹³Ca/¹³C β chemical shifts from random coil values were used to identify the type of secondary structure present in HET-s(218–289) fibrils. Negative deviations (Fig. 1f) indicated β -sheet secondary structure²⁰ for

¹The Salk Institute, 10010 North Torrey Pines Road, La Jolla, California 92037, USA. ²Laboratoire de Génétique Moléculaire des Champignons, Institut de Biochimie et de Génétique Cellulaires, Unité Mixte de Recherche 5095, Centre national de la Recherche Scientifique Université de Bordeaux 2, 33077 Bordeaux Cedex, France. ³ETH Zurich, Physical Chemistry, ETH Honggerberg, 8093 Zurich, Switzerland.

*These authors contributed equally to this work.

residues 226–234, 237–244, 262–271 and 273–282. Exceptions were observed at residues 229 and 265 (see below) and, to a lesser degree, at 277. Chemical shift assignment by solid-state NMR was achieved for all residues that showed slow hydrogen exchange (Fig. 1e). The striking correlation between the exchange data (Fig. 1e) and the chemical shift data (Fig. 1f) allowed for the confident establishment of secondary structure in the HET-s(218–289) amyloid fibrils: four β -strands comprising residues $\sim$ 226–234 (β 1), $\sim$ 237–245 (β 2),

$\sim$ 262–270 (β 3) and $\sim$ 273–282 (β 4) are connected by two short loops between β 1 and β 2, and β 3 and β 4, respectively, and by an unstructured, 15-residue-long segment between β 2 and β 3.

To determine the topology of β -strands in HET-s(218–289) fibrils, we made use of the architecture of β -sheets, in which the side chains of the even-numbered residues face in one direction, and those of the odd-numbered residues face in the opposite direction. Therefore, single cysteine residues were introduced at the odd- and

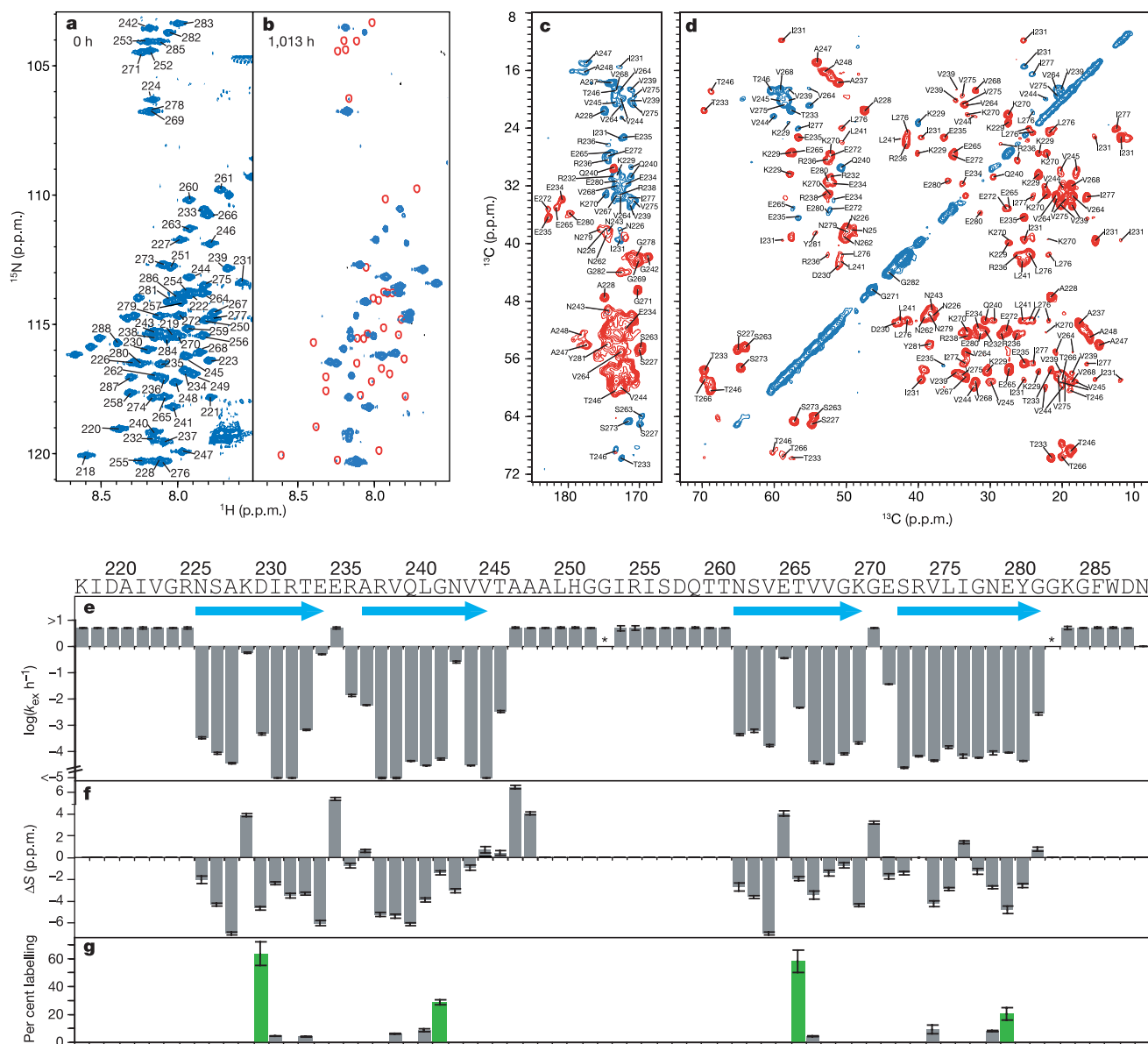


Figure 1 | Sequence-specific determination of regular secondary structure and its topology in HET-s(218–289) fibrils. **a, b**, Fast HMQC spectra of homogeneously ^{15}N -labelled HET-s(218–289) in d_6 -DMSO containing 0.1% d_1 -TFA, corresponding to fully protonated (**a**) and partially hydrogen exchanged (**b**) amyloid fibrils. Sequence-specific chemical shift assignments are labelled. Red lines encircle cross-peaks that show a virtually complete loss of intensity after $t_{\text{ex}} = 6$ weeks. **c, d**, Homonuclear ^{13}C - ^{15}N DREAM correlation spectra of ^{13}C , ^{15}N -labelled HET-s(218–289) fibrils. The carbonyl region of the DREAM spectra (**c**) was recorded at 40 kHz MAS. Positive contours (blue) were taken from a spectrum recorded with an up-down tangential DREAM sweep; negative contours (red) were taken from a similar spectrum but with a down-up sweep. Contour levels start at approximately 2.5 times root mean square (r.m.s.) noise level and increase by a factor of 1.4. Representative traces through the two-dimensional spectra are given in Supplementary Fig. S5. The aliphatic region of a DREAM spectrum (**d**) was

recorded at 25 kHz MAS. Sequence-specific chemical shift assignments are labelled. **e**, Plots of the observed quenched hydrogen exchange rates $k_{\text{ex}} \text{ h}^{-1}$. The measurable upper limit was 5 h^{-1} . Error bars indicate deviations from a monoexponential fit. Asterisks denote residues that could not be analysed, blue arrows the location of β -strands. **f**, Plot of $\Delta S = \Delta(\delta(^{13}\text{C}\alpha)) - \Delta(\delta(^{13}\text{C}\beta))$. $\Delta(\delta(^{13}\text{C}\alpha))$ and $\Delta(\delta(^{13}\text{C}\beta))$ are the difference between experimental $^{13}\text{C}\alpha$ and $^{13}\text{C}\beta$ chemical shifts and the corresponding 'random coil' chemical shifts. The line-width- and peak-overlap-dependent accuracy for each value is indicated. **g**, Solvent accessibility of single cysteine mutants. The fluorescence intensity of Alexa Fluor 488 crosslinked to the cysteine side chains is given relative to a positive control (see Methods) at the position at which the cysteine was introduced. Even-numbered positions are shown in green. Error bars indicate s.d. of at least three independent experiments.

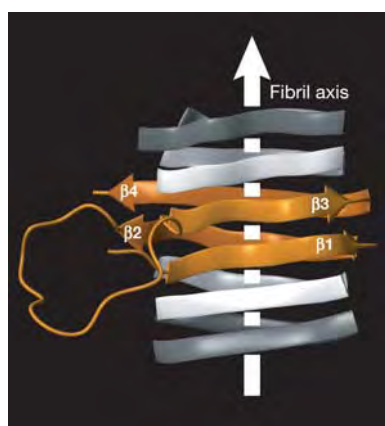


Figure 2 | The proposed fold of the infectious conformation of HET-s(218–289) amyloid fibrils. The extent and position in primary sequence of the β -strands is obtained from NMR data; the relative spatial position of the β -strands is modelled, taking into account the hydrophobicity of the residues, the high sequence identity between β -strands 1 and 3 and 2 and 4, and solvent accessibility data. There are, however, no direct inter-strand distance constraints. A model is shown as an orange ribbon diagram, flanked by neighbouring molecules indicated in white. β -strands are indicated by arrows, non-regular secondary structures by spine curves through the $C\alpha$ atoms of the corresponding residues, and the fibril axis by a white arrow. The β -strands are labelled.

even-numbered sides of each β -strand. Solvent accessibility was then probed by chemical crosslinking²¹ with a fluorescent dye using Alexa Fluor 488 C_5 maleimide. The extent of solvent accessibility of 11 Cys residues is shown in Fig. 1g. The even-numbered residues of each strand displayed high crosslinking rates, whereas all odd numbered

sides remained unlabelled. This indicated that all β -sheets have one solvent-accessible side.

The acquired structural data (Fig. 1) allowed us to propose a likely fold for the HET-s(218–289) fibrils (Fig. 2): the core of the fibrils consists of two β -strand-turn- β -strand motifs comprising β 1 and β 2 and β 3 and β 4, respectively, which are interconnected by a long loop formed by residues 246–261. The two segments interact through hydrogen bonding between β 1 and β 3, and β 2 and β 4, respectively, forming two layers of parallel β -sheets. The β -sheets continue along the fibril axis through intermolecular hydrogen bonding between neighbouring molecules (Fig. 2). The proposed fold is consistent with the NMR data, the observed pattern of solvent accessibility, the distribution of polar and hydrophobic residues on each strand, and the short loops between β 1 and β 2, and β 3 and β 4, respectively. It also takes into account the remarkable sequence identity of 28% between the segments β 1– β 2 and β 3– β 4, which suggests a repetitive arrangement (Supplementary Fig. S3). The only charged residues facing towards the inside of the fibril are K229 in β 1 and E265 in β 3, which are located in adjacent positions in the proposed fold, and therefore able to compensate their opposite charges. The reduced protection against hydrogen exchange of these residues as well as their non-beta solid-state chemical shifts (Fig. 1e, f) indicate a structural disturbance of the β -sheets at the positions of the proposed salt bridge.

In order to correlate the proposed fold of the HET-s(218–289) fibrils with [Het-s] function and infectivity, a mutational approach was used. Because the fold is based on β -sheet secondary structure, we introduced proline residues, which are known to act as local β -strand breakers. A series of 13 mutants was generated with a single proline substitution approximately every five residues in the 218–289 region of full-length HET-s (Fig. 3a). A number of short deletion mutants affecting the central loop and the β -strand segments were also generated (Fig. 3b). *P. anserina* transformants expressing the

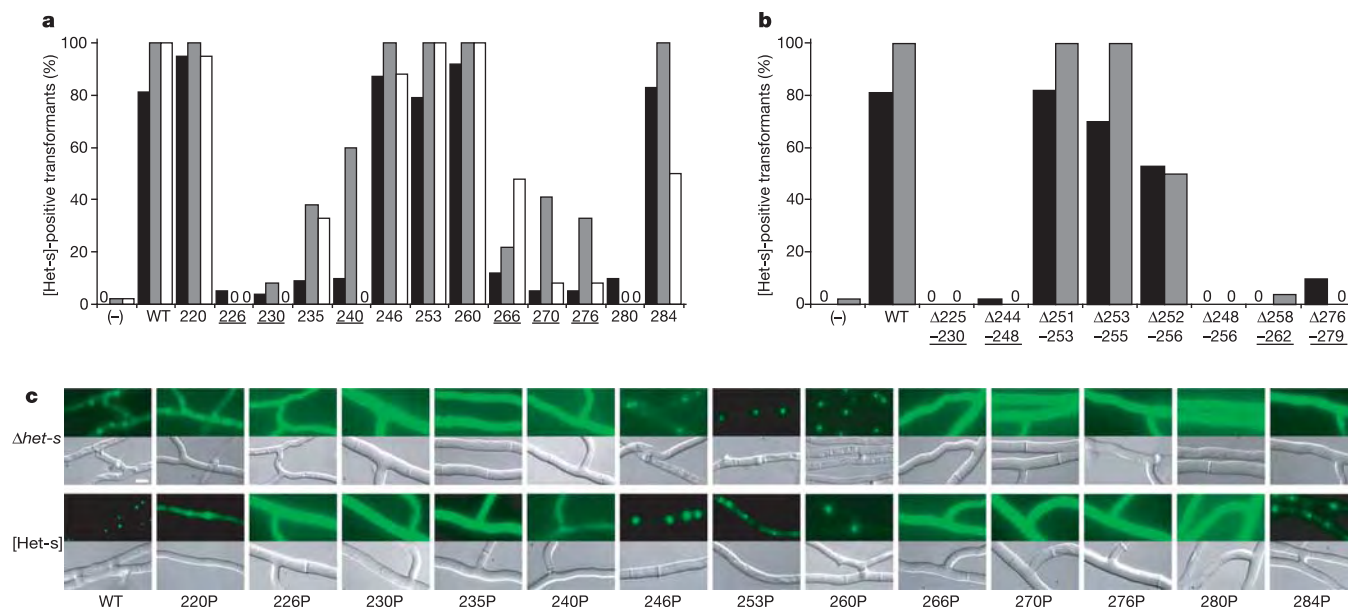


Figure 3 | In vivo prion formation of HET-s proline and deletion mutants.

a, Proline mutants were introduced into a Δ het-s *P. anserina* strain and the transformants prion-infected by confrontation with a [Het-s] colony. Given are the percentage of transformants producing a cell death reaction when confronted with a het-S tester colony (black columns), and the percentage of transformants displaying [Het-s] infectivity immediately after infection (grey columns) and after a 3-day subculture (white columns). The minus sign indicates vector alone, and WT indicates wild-type HET-s; mutants are labelled by the position of the proline. Zero indicates that all transformants were negative. **b**, Deletion mutants of HET-s were expressed in a Δ het-s

strain and the transformants prion-infected as above. Colour code as in **a**.

c, In vivo aggregation of HET-s-GFP fusion proteins with proline substitutions. Wild-type and mutant HET-s-GFP fusion proteins were expressed either in a Δ het-s knockout strain or a wild-type prion-infected [Het-s] strain. Transformants of the Δ het-s background were prion-infected by contact with a wild-type [Het-s] strain before microscopic observation. Scale bar, 4 μ m. Mutants located in β -strand elements are underlined. A complete data set with the numbers of individual transformants tested is given in Supplementary Table 1.

mutant proteins were prion-infected by contact with a wild-type [Het-s] colony. The presence of [Het-s] prions in these transformants was then assayed by monitoring their ability to produce a cell death reaction when fused to a HET-S-expressing colony, and to infect a prion-free [Het-s*] colony. [Het-s] infectivity was assayed immediately after contact of the transformants with the wild-type [Het-s] donor and again after a 3-day subculture (Fig. 3a; see also Supplementary Table 1). In order to monitor the formation of HET-s aggregates *in vivo*, the same mutations were introduced in a HET-s–GFP fusion protein. The mutant fusion proteins were over-expressed both in *P. anserina* Δ het-s knockout strains and in a [Het-s] strain in order to determine whether co-expression with endogenous wild-type prions would lead to aggregation of the mutant fusion proteins (Fig. 3c).

In striking correlation with the structural data, proline substitutions located in the β -strand regions strongly affected [Het-s] function, infectivity (Fig. 3a) and HET-s–GFP foci formation (Fig. 3c). In contrast, proline substitutions in the N- or C-terminal flexible tails or the central loop did not affect [Het-s] activity, infectivity, or HET-s–GFP foci formation. Similarly, the deletions Δ 225–230, Δ 244–248, Δ 258–262 and Δ 276–279, which affected the β -strand elements, abolished [Het-s] infectivity. Short deletions of no more than three amino acids in the central loop did not affect [Het-s], but a larger deletion, Δ 248–256, led to loss of [Het-s] function and infectivity. This indicates that a minimal length in the central loop might be required.

For some mutations in β -strand positions, [Het-s] prion infectivity could be detected transiently, but was lost upon subculturing (for instance, HET-s 240P, Fig. 3a). Concomitantly, GFP foci formation of HET-s 240P was detected only when the mutant protein was co-expressed with wild-type [Het-s] prions (Fig. 3b). This suggests that these mutant proteins retained a certain ability to form [Het-s] prions, but that they are mitotically very unstable. In addition, a number of proline substitutions in β -strand regions and in the short connecting loop did not lead to a total loss of [Het-s] function or infectivity, indicating that they interfered only partially with amyloid formation. Indeed, recombinant HET-s 226P and 266P retained a residual ability to form amyloid fibrils *in vitro* and to infect prion-free [Het-s*] colonies when introduced biolistically (Supplementary Fig. S4 and Supplementary Table 2). Interestingly, HET-s 246P retained [Het-s] activity. Residue 246 is still involved in hydrogen bonding, but has an α -helical rather than a β -sheet chemical shift. This suggests that it is specifically the β -sheet secondary structure that is required for HET-s function and infectivity.

Taken together, the local disruption of any of the four β -strands reduced or even abolished the formation and function of the [Het-s] prion, whereas the central loop only required a certain minimal length to sustain prion formation.

Our data correlate strongly the amyloid conformation with infectivity. Hence, the proposed fold of the HET-s(218–289) fibrils shown in Fig. 2 represents an infectious protein conformation. Because the [Het-s] prion has a natural function in heterokaryon incompatibility, it seems plausible that this infectious conformation was evolutionarily optimized for amyloid fibril formation. This view is supported by the extraordinarily well-ordered structure as manifested in the narrow line width observed in the solid-state NMR spectra, and the monoexponential hydrogen exchange rates. In contrast, other amyloid fibrils associated with misfolding diseases appear to be structurally more heterogeneous^{22–25}. Furthermore, the dimer-like double β -strand-turn- β -strand motif formed by the pseudo-repeats β 1– β 2 and β 3– β 4 may have evolved to optimise the polymerization: in the case of a single β -strand-turn- β -strand motif, the initial step of fibril growth requires the diffusion-dependant formation of an oligomeric nucleus²⁶. The covalent nature of the pseudo-repeats is therefore likely to promote this nucleation event by reducing its diffusion dependency.

It is generally believed that most prions will share β -sheet-rich

amyloid fibrils as a common structural feature. Here, we have now shown that a β -sheet structure is indeed the infectious prion conformation of HET-s. This link between structure and infectivity constitutes a significant step towards the elucidation of the mechanism of prion formation and propagation.

METHODS

Mutagenesis. All site-directed mutants were generated with the Quick-change kit (Stratagene) using the pCB1004-het-s, pGPD-het-s–GFP¹⁴ and the pET-21a-het-s vectors as template¹².

Preparation of stable-isotope-labelled HET-s(218–289) fibrils. Stable-isotope-labelled HET-s(218–289) was expressed as described for other HET-s constructs¹³. The bacterial pellet was solubilized in 50 mM TRIS pH 8.0 and 150 mM sodium chloride (TCL) containing 6 M guanidinium hydrochloride. The supernatant was centrifuged for 30 min at 18,000g. The protein was purified from the supernatant by His₆-affinity chromatography and concentrated to approximately 0.5–1 mM. Fast buffer exchange to TCL buffer yielded monomeric HET-s(218–289), which started to aggregate immediately into amyloid fibrils.

Hydrogen exchange. ¹⁵N-labelled HET-s(218–289) was used. To reduce the size of large clumps of fibrils, a suspension of the fibrils in liquid nitrogen was bruised with a mortar. To start the hydrogen exchange, the fibrils were sedimented at 3,600g for 10 min, washed with 50 mM TRIS pH 8.0^{read} containing 150 mM sodium chloride, 5 mM dithiothreitol and D₂O as the solvent, and re-suspended in the same buffer. At suitable intervals, aliquots were sedimented at 24,000g for 2 min, washed with D₂O and frozen in liquid nitrogen to quench hydrogen exchange. For the NMR analysis, the fibrils were dissociated in perdeuterated dimethylsulphoxide (d₆-DMSO) containing 0.1% deuterated trifluoroacetic acid (d₁-TFA). The amount of residual D₂O was about 3%. Immediately after dissolving the fibrils, a series of 80 two-dimensional [¹⁵N, ¹H] correlation spectra were measured for 8 h. To distinguish residues that display fast exchange in the fibrils and residues that have high intrinsic exchange rates in DMSO, which would both result in absent peaks in the [¹⁵N, ¹H] correlation spectrum¹⁷, a second series of 80 two-dimensional spectra were measured upon addition of 3% H₂O. Using this control measurement, residues 253 and 283 were excluded from the analysis. With the exception of the amide group of N289, a complete sequence-specific assignment of the backbone H^N/N cross-peaks was obtained using the triple-resonance experiments HNCA²⁷ and HNCA(CO)NH²⁸ applied to ¹³C, ¹⁵N-labelled HET-s(218–289). The data were analysed using the programs PROSA²⁹ and CARA (<http://www.nmr.ch>), and a specially written Visual basic program in combination with Microsoft Excel¹⁷.

Solid-state NMR. The HET-s(218–289) fibrils were washed in H₂O and centrifuged into MAS rotors sealed with two-component epoxy adhesive to prevent dehydration. Solid-state NMR spectra were recorded on a Bruker AV600 spectrometer at a static field of 14.09 T. A 2.5-mm Chemagnetics probe was used for the experiments at 25 kHz MAS and a 1.8-mm probe constructed by A. Samoson³⁰ for the experiments at 40 kHz MAS. The sample temperature was kept at 5–15 °C. For the ¹³C–¹³C correlation spectra recorded using the DREAM scheme¹⁸, the contact time of the initial adiabatic cross-polarization was 1 ms and the RF-field strengths were 70 kHz and 90 kHz for ¹³C and ¹H at 25 kHz MAS, and 90 kHz and 130 kHz for ¹³C and ¹H at 40 kHz MAS, respectively. Continuous wave decoupling was applied during the DREAM mixing step and XiX decoupling during *t*₁ and *t*₂, both with an RF-field amplitude of 150 kHz. The measurement time was 33 h and 55 h for the experiments at 25 kHz and 40 kHz MAS, respectively. These spectra we used for the identification of spin systems and the ¹³C side-chain assignment. For the backbone assignment we recorded the following heteronuclear two-dimensional correlation spectra: NCA, NCO, N(CA)CO, N(CA)CB and N(CO)CA. The assignment was done using the CARA program (<http://www.nmr.ch>) and confirmed via a CA–CA correlation spectrum.

Side-chain solvent accessibility studies. Proteins were purified as described above. Fibrils were reduced with 20 mM dithiothreitol before the final buffer exchange, and re-natured under nitrogen. To label solvent-accessible Cys residues by chemical crosslinking, 85 μ M fibrils were incubated with a 2.5-fold molar excess of Alexa Fluor 488 C₅ maleimide (Molecular Probes) for 20 min at room temperature. The fibrils were washed extensively with TCL buffer and solubilized with 8 M spectroscopically pure guanidine hydrochloride. For fluorometry, the samples were diluted at least 70-fold into TCL buffer. The Alexa Fluor 488 fluorescence (excitation 493 nm, emission 516 nm) was measured relative to the fluorescence of a single Trp residue in HET-s(218–289) (excitation 295 nm, emission 353 nm). To determine the maximum labelling efficiency, the mutant HET-s(218–289) D288C was used, which is located in the presumably solvent accessible C-terminal region showing fast

hydrogen exchange. Wild-type HET-s(218–389) fibrils were labelled to $1.9 \pm 0.1\%$.

In vivo analysis of HET-s mutants. *het-s* mutant alleles were introduced in the Δ *het-s* and wild-type *het-s* recipient strains¹⁰. [Het-s] incompatibility function, [Het-s] infectivity and HET-s–GFP foci formation were analysed as described previously¹⁵. After their initial contact with the [Het-s] prion donor, the transformants were either subcultured for 3 days or directly confronted with the [Het-s*] recipient in order to be able to detect transient expression of the [Het-s] state. Under the tested conditions, spontaneous emergence of [Het-s] in [Het-s*] colonies is in the range of 2%.

Received 4 March; accepted 12 May 2005.

- Alper, T., Cramp, W. A., Haig, D. A. & Clarke, M. C. Does the agent of scrapie replicate without nucleic acid? *Nature* **214**, 764–766 (1967).
- Prusiner, S. B. Novel proteinaceous infectious particles cause scrapie. *Science* **216**, 136–144 (1982).
- Sparrer, H. E., Santos, A., Szoka, F. C. Jr & Weissman, J. S. Evidence for the prion hypothesis: induction of the yeast [PSI⁺] factor by *in vitro*-converted Sup35 protein. *Science* **289**, 595–599 (2000).
- King, C. Y. & Diaz-Avalos, R. Protein-only transmission of three yeast prion strains. *Nature* **428**, 319–323 (2004).
- Tanaka, M., Chien, P., Naber, N., Cooke, R. & Weissman, J. S. Conformational variations in an infectious protein determine prion strain differences. *Nature* **428**, 323–328 (2004).
- Maddelein, M. L., Dos Reis, S., Duvezin-Caubet, S., Couлары-Salin, B. & Saupe, S. J. Amyloid aggregates of the HET-s prion protein are infectious. *Proc. Natl Acad. Sci. USA* **99**, 7402–7407 (2002).
- Legname, G. *et al.* Synthetic mammalian prions. *Science* **305**, 673–676 (2004).
- Glass, N. L. & Kaneko, I. Fatal attraction: nonself recognition and heterokaryon incompatibility in filamentous fungi. *Eukaryot. Cell* **2**, 1–8 (2003).
- Saupe, S. J. Molecular genetics of heterokaryon incompatibility in filamentous ascomycetes. *Microbiol. Mol. Biol. Rev.* **64**, 489–502 (2000).
- Turcq, B., Deleu, C., Denayrolles, M. & Begueret, J. Two allelic genes responsible for vegetative incompatibility in the fungus *Podospora anserina* are not essential for cell viability. *Mol. Gen. Genet.* **228**, 265–269 (1991).
- Coustou, V., Deleu, C., Saupe, S. & Begueret, J. The protein product of the *het-s* heterokaryon incompatibility gene of the fungus *Podospora anserina* behaves as a prion analog. *Proc. Natl Acad. Sci. USA* **94**, 9773–9778 (1997).
- Dos Reis, S. *et al.* The HET-s prion protein of the filamentous fungus *Podospora anserina* aggregates *in vitro* into amyloid-like fibrils. *J. Biol. Chem.* **277**, 5703–5706 (2002).
- Balguerie, A. *et al.* Domain organization and structure-function relationship of the HET-s prion protein of *Podospora anserina*. *EMBO J.* **22**, 2071–2081 (2003).
- Coustou-Linares, V., Maddelein, M. L., Begueret, J. & Saupe, S. J. *In vivo* aggregation of the HET-s prion protein of the fungus *Podospora anserina*. *Mol. Microbiol.* **42**, 1325–1335 (2001).
- Balguerie, A. *et al.* The sequences appended to the amyloid core region of the HET-s prion protein determine higher-order aggregate organization *in vivo*. *J. Cell Sci.* **117**, 2599–2610 (2004).
- Hoshino, M. *et al.* Mapping the core of the β 2-microglobulin amyloid fibril by H/D exchange. *Nature Struct. Biol.* **9**, 332–336 (2002).
- Lührs, T. *et al.* The 3D structure of Alzheimer's A β (1–42) fibrils. *Nature* (submitted).
- Verel, R., Ernst, M. & Meier, B. H. Adiabatic dipolar recoupling in solid-state NMR: The DREAM scheme. *J. Magn. Reson.* **150**, 81–99 (2001).
- Siemer, A. B., Ritter, C., Ernst, M., Riek, R. & Meier, B. H. High-resolution solid-state NMR of the prion protein HET-s in its amyloid conformation. *Angew. Chem. Int. Edn Engl.* **44**, 2441–2444 (2005).
- Wishart, D. S. & Sykes, B. D. The ¹³C chemical-shift index: a simple method for the identification of protein secondary structure using ¹³C chemical-shift data. *J. Biomol. NMR* **4**, 171–180 (1994).
- Javitch, J. A., Shi, L. & Liapakis, G. Use of the substituted cysteine accessibility method to study the structure and function of G protein-coupled receptors. *Methods Enzymol.* **343**, 137–156 (2002).
- Tycko, R. Progress towards a molecular-level structural understanding of amyloid fibrils. *Curr. Opin. Struct. Biol.* **14**, 96–103 (2004).
- Petkova, A. T. *et al.* Self-propagating, molecular-level polymorphism in Alzheimer's β -amyloid fibrils. *Science* **307**, 262–265 (2005).
- Laws, D. D. *et al.* Solid-state NMR studies of the secondary structure of a mutant prion protein fragment of 55 residues that induces neurodegeneration. *Proc. Natl Acad. Sci. USA* **98**, 11686–11690 (2001).
- Yamaguchi, K. *et al.* Core and heterogeneity of β 2-microglobulin amyloid fibrils as revealed by H/D exchange. *J. Mol. Biol.* **338**, 559–571 (2004).
- Harper, J. D. & Lansbury, P. T. Jr Models of amyloid seeding in Alzheimer's disease and scrapie: mechanistic truths and physiological consequences of the time-dependent solubility of amyloid proteins. *Annu. Rev. Biochem.* **66**, 385–407 (1997).
- Grzesiek, S. *et al.* ¹H, ¹³C, and ¹⁵N NMR backbone assignments and secondary structure of human interferon- γ . *Biochemistry* **31**, 8180–8190 (1992).
- Bracken, C., Palmer, A. G. III & Cavanagh, J. (H)N(COCA)NH and HN(COCA)NH experiments for ¹H–¹⁵N backbone assignments in ¹³C/¹⁵N-labeled proteins. *J. Biomol. NMR* **9**, 94–100 (1997).
- Guntert, P., Dotsch, V., Wider, G. & Wuthrich, K. Processing of multidimensional NMR data with the new software Prosa. *J. Biomol. NMR* **2**, 619–629 (1992).
- Samoson, A., Tüherm, T. & Past, J. Rotation sweep NMR. *Chem. Phys. Lett.* **365**, 292–299 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements R.R. is a Pew scholar. This research was supported in part by grants from the National Institute of Health, the US Army, the ETH Zurich, the Swiss National Science Foundation, the CNRS and the French Ministry of Research.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.R. (riek@salk.edu).

-  FOCUS
-  SPOTLIGHT
-  RECRUITMENT
-  ANNOUNCEMENTS
-  EVENTS

naturejobs

French lessons

It is relatively common for a young French PhD to do a fellowship in the United States — but it is pretty rare for a US-born PhD to undertake a postdoc in France. This unequal balance could soon be evened out thanks to a scheme dreamt up and organized by BioTeam Paris (www.bioteam-parisregion.org).

BioTeam has set up a fellowship aimed at teaming an institution or company in North America with another in Paris. The winning entry will receive funding to place a North American postdoc with a Paris-based company — thus cementing the fledgling partnership.

The selection process for this year's inaugural award bridged the geographical and cultural divides by convening a committee with members from academia, industry and media hailing from both sides of the Atlantic (a group that this year included myself). Like all study sections, the first criterion the applicants were judged on was the strength of their science. But the remit also asked for a strong partnership, based on complementary expertise, intellectual property and marketing strengths.

Although all five short-listed applicants met those

criteria, an entry by US-based technology company Agilent and Paris-based ExonHit Therapeutics stood out. Agilent's expertise in developing gene- and protein-based chips seemed to complement ExonHit's interest in alternative RNA splicing — perhaps an alliance between the two could help develop an alternative splice chip? And, especially important to me, the potential partners defined a clear R&D role for their possible postdoc — not just as another pair of hands.

This award, presented last week in Paris, could do more than bring one North American postdoc across the Atlantic for fine wine and excellent cuisine. It could serve as an example of how to develop international partnerships where every entity benefits. And give North American researchers a reason to work abroad.



Paul Smaglik, Naturejobs editor

CONTACTS

Publisher: Ben Crowe
Editor: Paul Smaglik
Marketing Manager: David Bowen

US Head Office, New York
 345 Park Avenue South, 10th Floor,
 New York, NY 10010-1707
 Tel: +1 800 989 7718
 Fax: +1 800 989 7103
 e-mail: naturejobs@natureny.com

US Sales Manager/Corporations:
 Peter Bless
 Classified Sales Representatives
 Tel: +1 800 989 7718

**New York/Pennsylvania/
 Latin America:** Kelly Roman
**Midwest USA/Maryland/
 NIH:** Wade Tucker
East USA/Canada:
 Janine Taormina

San Francisco Office
Classified Sales Representative:
 Michaela Bjorkman
 West USA/West Corp. Canada
 225 Bush Street, Suite 1453
 San Francisco, CA 94104
 Tel: +1 415 781 3803
 Fax: +1 415 781 3805
 e-mail: m.bjorkman@naturesf.com

European Head Office, London
 The Macmillan Building,
 4 Crinan Street,
 London N1 9XW, UK
 Tel: +44 (0) 20 7843 4961
 Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com

Naturejobs Sales Director: Nevin Bayoumi (4978)
European Sales Manager: Andy Douglas (4975)

Advertising Production Manager: Billie Franklin
 To send materials use London address above.
 Tel: +44 (0) 20 7843 4814
 Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com

Naturejobs web development: Tom Hancock
Naturejobs online production: Niamh Shields

European Satellite Office
**Germany/Austria/Belgium/
 The Netherlands:**
 Patrick Phelan, e-mail: p.phelan@nature.com
 Reya Silao, e-mail: r.silao@nature.com

Japan Head Office, Tokyo
 MG Ichigaya Building (5F), 19-1 Haraikatamachi,
 Shinjuku-ku, Tokyo 162-0841
 Tel: +81 3 3267 8751
 Fax: +81 3 3267 8746
Asia-Pacific Sales Director: Rinoko Asami
 e-mail: rasami@naturejpn.com

Getting schooled

When Carol Thornber took her first faculty position last autumn, she inherited teaching responsibility for two undergraduate courses. Although Thornber, a marine ecologist, had to some extent been prepared by a unique postdoctoral training programme, making the transition from student to lecturer was far from plain sailing.

Her autumn course in upper-level marine botany, at the University of Rhode Island in Kingston, had her scrambling to learn East Coast species of algae and scouting field-trip sites. Her spring class, introductory biology for some 130 first-year undergraduates, had her struggling with endless administrative duties. But by the middle of the year, Thornber says, she had made a discovery that changed the way she worked.

"I realized I couldn't let teaching take all of my time," says Thornber. "To get tenure here, I need to do research and I like to do research. So I took one day a week and kept it as a research day."

Teaching, especially undergraduate courses, with all its preparation and in-class time, takes up more of a young lab head's time than anything else except research duties. Unfortunately, it is often an activity in which a rookie investigator has very little experience. Facing a classroom of eager young minds may be overwhelming, particularly for junior faculty members who have never led a course themselves.

But there are some techniques you can learn to save yourself from drowning. Some fellowships and programmes exist to prepare researchers to teach (see 'That'll teach you', opposite). But most new investigators



Stars of tomorrow: many lab heads are moving away from straight lectures to teaching systems, such as putting students into 'learning teams'.

won't get such formal training. Instead, they should rely on mentors and scientific societies for resources. New teachers can also develop strategies, in terms of organizing materials and starting small. Once you've cracked the time-management dilemma, you can innovate and adjust — and ideally create both a balance and an interplay between your research and teaching.

Once in a junior faculty position, all veterans agree on the first step: find a teaching mentor. Ideally, a mentor should be someone who pays attention to the educational developments in your field of science and whose teaching style you admire.

You're not the first

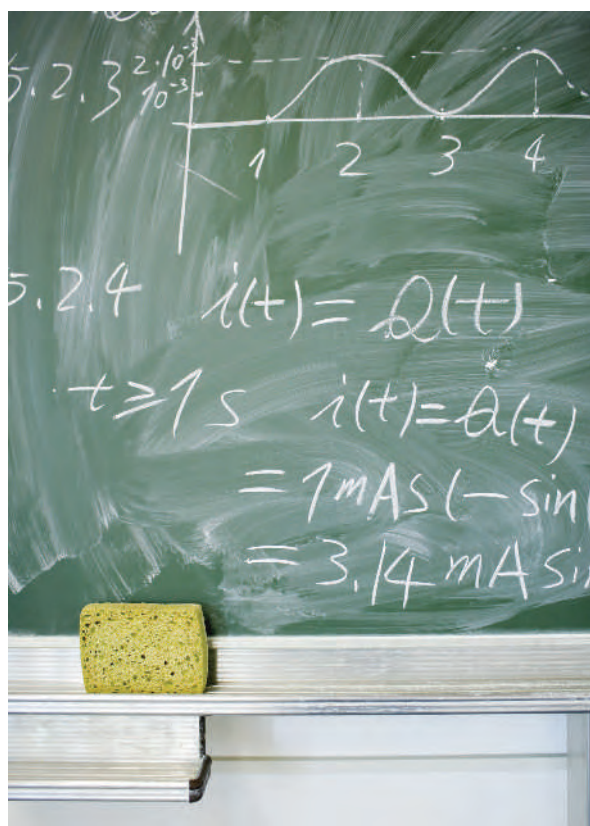
"Have someone be your guide who's already made the mistakes you don't have to make," says Mike Zeilik, an emeritus professor of astronomy at the University of New Mexico in Albuquerque. No one is the first to teach any type of class, no matter how much it feels that way. Zeilik's own mentor was an engineer who invited him to tag along to a meeting on physics education.

That experience and his later work on improving astronomy education opened his eyes to another important source of help for new investigators. "Most professional societies have a committee on teaching or education. Find out who those people are and get in touch to find out what resources are available," he suggests. There are myriad tools for typical introductory courses in almost every field, he says, and these can sometimes be found at the committees' websites.

More and more science departments are turning away from the standard straight lecture format. Zeilik and other experts recommend keeping up with the best-tested practices for active cooperative learning — this could make your teaching job easier and more effective in the long-run.

To juggle teaching and research, classroom veterans advise beginners to draw on mentors and all the resources they can find.

Kendall Powell learns about the balancing act.



For her introductory biology course, Thornber relied heavily on the prepared materials and slides that came with the textbook. She also took advantage of the extra administrative support her department provides for large courses. But even in the algae course — her specialist subject — Thornber realized she needed help from more experienced colleagues.

Start small if you have the chance, says Japanese native Masayuki Numata, who set up his molecular cell biology lab at the University of British Columbia in Vancouver three years ago. As English isn't his first language, Numata opted to ease into his teaching responsibilities with a very small course tutoring medical students. His major duty was to stimulate discussion among students dissecting mock patient case studies.

"In the future, I'll have to teach bigger classes, but now I can handle it with less stress," says Numata. "Now I know how to encourage students to learn and have established my teaching style."

In at the deep end

Not everyone has the luxury of a such a transition. Although some institutions give new researchers a teaching hiatus of one or two years, others expect you to jump right in. No matter where you land, says Jerry Hedrick, director of the Professors of the Future programme at the University of California, Davis, "there will be teaching-intensive times and research-intensive times". Plan accordingly by looking at your two-year and five-year goals and — just like a grant deadline — give yourself a cut-off point for teaching preparation, he advises.

Darcy Kelley, a neurobiologist at Columbia University in New York, agrees. "Time management is of the essence for junior faculty members," she says. "Use the law of diminishing returns and decide what's necessary and what's obsessing." One trick she uses is to teach all of her courses in one semester, to ensure that her autumn semester is free to focus solely on moving research projects forward. Also, she notes that teaching afternoon classes forces her to spend only the morning hours that day to prepare.

Both new and experienced teachers emphasize the importance of time management, and warn beginners not to waste effort on 'reinventing the wheel'.

Thornber follows wisdom she found in Robert Boice's *Advice for New Faculty Members* (Allyn and Bacon, Needham Heights, Massachusetts, 2000). She begins preparing for lectures early by hammering out a quick outline. "When I sit down to write a lecture



Making a point: Carol Thornber, right, quickly realized that she had to strike a balance between teaching duties and research.

I already have some notes and don't have a mad crunch right at the end."

Mentally separating your commitments can help lessen stress, too. "It doesn't help if you are teaching and thinking, 'I should be writing a paper now,'" says Susanne Mandrup, a molecular biologist at the University of Southern Denmark, Odense. "I often think I have too many other things to do, but once I'm in the classroom, it's fun. It's better to be positive about it and take the students seriously," she advises.

"Be prepared" is the ultimate time-management advice from Eric Mazur, an applied physicist and education researcher at Harvard University. "Doing a good job teaching is not necessarily more time-consuming than doing a bad job of teaching," he notes. Standing in front of 500 first-years is not that different from standing in front of 500 colleagues at a meeting, he says — and is much less intimidating.

Think of it as an art form

In fact, once you have the organizational and time-management aspects cracked, it's time to think about teaching as experiment or art form. "Do not hold back in innovating in your own teaching," advises Zeilik. "Experiment, you are an experimentalist." The trick, he says, is that as in any experiment, you have to assess the outcome and revise the innovation accordingly.

Others who advocate co-teaching with more experienced faculty members point out that science is never done alone, so why should teaching be done solo?

"Part of doing research is knowing how to convey your results, the methods, and excitement of science to other people," says Mazur. Lecturers are not only preparing the next generation of scientists, he says, but polishing their own communication skills and strengthening public support of scientific research.

Indeed, established researchers say that approaching your teaching in the same manner you approach your research will make it more valuable to you and your students. Instruction, far from being just a necessary distraction from the bench, they say, should be viewed as another part of the job that can be integrated with your research goals.

Kendall Powell is a freelance writer based in Broomfield, Colorado.

WEB LINKS

NIGMS career development awards
www.nigms.nih.gov/minority/iracda_institutions.html
 Professors of the Future
prof.ucdavis.edu

THAT'LL TEACH YOU

The ideal situation for would-be junior faculty members is to get teaching experience under their belts before they land their first job. The Professors of the Future (PROF) programme at the University of California, Davis, is one such programme. It teams postdocs up with professors from nearby San Francisco State University to co-teach courses. The programme is one of seven career-development programmes sponsored by the National Institute of General Medical Sciences in Bethesda, Maryland. Each programme brings together a research-intensive university and at least one minority-serving university or community college.

Matching future teachers to institutions is as important as matching your own goals to your training, says Jerry Hedrick, PROF director. Job candidates should know what they are signing up for when taking a job at teaching-intensive or research-intensive institutions, he says. Matching a department's expectations and objectives to your own will help avoid frustrations later.

K.P.

Looking for Mr Goodbug

Will work for food.

Elisabeth Malartre

It all began with bedbugs. One night at 'Paradise Cabins' in Eureka left Jim Chiampi with red splotches sprinkled all over his body. For the next few itchy days in his lab at UCI, he toyed with how to fix bedbugs so that they wouldn't find humans attractive.

Seymour Benzer's work on the heritability of behaviour in fruitflies started people looking for behaviour genes. Is there a 'gay' gene? A gene for aggressiveness? What about the 'depression' gene?

The cloak of culture wraps human behaviour; it's difficult to sort out learned from instinctive behaviour. Not so with insects. Parasitic male wasps hatch from a moth egg ready to mate within minutes. Virgin females are already interested in smooth egg-shaped objects. No learning there.

Chiampi idly watched ants on the lab windowsill. They were coming and going from the potted orchid: living in it, but not harming it. Images of army ants flashed in his mind, a parade of millions devouring the countryside, waving bits of leaves. Could he make these ants attack and destroy their orchid?

He vaguely remembered an article in *Nature*. Leafcutter ants (*Atta* spp.) were mainly tropical, one species just recently sequenced. But importing leafcutters was impossible. He'd need to modify a native ant species. Nothing too small; a head size of two millimetres is ideal for leafcutters. He settled on local harvester ants. Primarily seed-eaters, they dismantle vegetation when it suits them, surrounding their mounds with shredded plant bits.

By 2015, cutting and pasting genes into animals was routine. Chiampi consulted the Genes2Go digital catalogue: he downloaded the genome for the one *Atta* species and several non-leafcutters for comparison. After weeks of looking, he identified the leafcutting genes and fed the sequences into his synthesizer.

Two hurdles to overcome: programming the local ants to shred plants for food instead of collecting seeds, then getting them to find specific plants among non-target species.

After several hilarious failures, and some painful bites from frustrated genetically modified ants, he was ready to field-test strain HA-12.

He and a few friends built a large, hermetically sealed, domed sandbox and

embedded a small pampas-grass plant in it. Then he mated the neck of the ant colony box to the enclosure and lifted a tiny portcullis gate to let the ants out.

Slowly, a few scouts appeared and spread out over the sand. One found the pampas grass and began waving its antennae. It headed to the nest, leaving a scent trail back to the plant. Forager ants emerged from the colony, like tiny cows heading to pasture.

Chiampi squatted next to the nest box and popped open a beer. Within minutes, the ant trail broadened from single file to several abreast, pouring out of the colony until a solid line of ants marched towards the doomed plant. Soon they were carrying pieces of plant back towards the nest. Watching them was hypnotic.

"How long to take down that plant?" asked his sister Marion.

"An average colony of leaf cutters can eat as much plant material in a day as a cow. I've got only a few hundred ants in there. I figure maybe an hour."

The ants beat that easily.

A sceptical friend recorded the test on a cell phone and posted it on the Internet. Millions watched as the ants swarmed, the plant shivered, then collapsed.

Herbicide manufacturers protested, sending squads of 'concerned citizens' to Congress to protest about the dangers of genetically modified ants. Pickets wearing huge mutated fly suits and carrying "Frankenbugs no!" signs milled outside Chiampi's lab. The news vultures were enthralled — waiting, expecting the weight of outraged public opinion to crush GM ants, as it had crippled GM foods 20 years before.

But the public, sick of battling ants in the kitchen and dandelions in the lawn, loved the idea. Turnabout is fair play. If we can't stop 'em, make 'em useful.

Chiampi hurriedly switched his ants to eating dandelions. He called his start-up 'Dandi-Ants' and moved out of his university lab and into a tilt-up.

Targeting was easy. Chiampi sprayed some target plants with a pheromone (the "scent of death", Marion called it)



to attract the scouts. After a few bites, the taste of the plant itself kept the ants focused. Then they cut up all the others like it.

Chiampi's research jump-started the entire formi-farming industry. Beekeepers once trucked their hives to orchards for pollination services. Now formi-farmers offer mobile weed eradication from ant colony trucks.

Ants go everywhere; no cliff is too steep, no perch too isolated. They're highly motivated workers, harvesting their food. And by keeping the queens securely in the colony, the industry ensures that the sterile workers return to the nest at night.

On California's steep hillsides ants now dutifully clear fuel-modification zones around houses, replacing expensive prison labour or damaging herds of goats, while leaving sensitive plant species untouched. For especially common weeds, there are specialized strains of ants.

Today, the homeowner can hire legions of ants. Crabgrass in the lawn? Mint overrunning the petunias? No problem! We'll be over in an hour. No more weekends spent digging in the dirt! And no more poisons!

The kudzu weed has stopped eating Georgia. The gorgeous Big Sur coastline has emerged from a curtain of pampas grass. In the western deserts, ants eliminated water-sucking tamarisk. Springs reappeared. Wildlife flourished.

Humans are learning to harness the little things that run the world, to help save it.

Elisabeth Malartre is the pen name of a biologist and writer living in Laguna Beach, California, who teaches at the University of California, Irvine.

JACEY